DOCUMENT RESUME

ED 337 480 TM 017 294

TITLE	Proceedings of the 1983 IPMAAC Conference on Public Personnel Assessment (7th, Washington, D.C., May 22-26, 1983).
INSTITUTION	International Personnel Management Association, Washington, DC.
PUB DATE	May 83
NOTE	149p.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Assessment Centers (Personnel); Computer Assisted Testing; *Evaluation Methods; Job Analysis; *Job Performance; *Occupational Tests; *Personnel Evaluation; Personnel Management; Personnel Selection; Predictive Measurement; *Public Sector; *Test Use
IDENTIFIERS	International Personnel Management Association

ABSTRACT

The International Personnel Management Association Assessment Council (IPMAAC) focuses on improving public personnel assessment in such fields as selection and performance evaluation. Author-generated summaries/outlines of papers presented at the IPMAAC's 1983 conference are provided. The presidential address is "Is There a Future for Test Writers?" by C. B. Schultz. The keynote address is "Policies and Issues Facing EEOC (Equal Employment Opportunity Commission) Today" by A. Golub. The following paper sessions are reviewed: "Job Analysis Methodologies and Applications"; "An Introduction to Assessment Centers: Development and Uses"; "Evaluating Other Minorities: The Handicapped and the Elderly"; "Behavioral Traits and Their Influence on Selection Procedures"; "Assessment Centers: Research and Application"; "The Effects of Recession, Labor Relations, and Reform of Public Agencies"; "New Developments in Selection for the Fire Service"; "Biodata as a Public Sector Selection Strategy"; "Focus on the Economics of Testing: Measuring Utility"; "Training and Experience Evaluations and Other Forms of Self-Assessment"; "Exploring Issues Related to Fairness, Adverse Impact, and Test Bias"; "Assessing Managerial Skills"; "Alternatives and Traditional Selection Procedures"; "Measuring Productivity"; "Physical Ability Testing"; and "The Development and Validation of Selection Standards for Law Enforcement". The following symposia are reviewed: "Reasonable Accommodation in Selection and Employment of Deaf White-Collar Employees"; "Development of a Valid Selection Procedure for Nineteen Professional Classes in a State"; "Development of Job Related Valid Oral Board Examinations"; "Firefighter Selection: Non-Written Tests"; "Designing and Implementing Personnel Functions in the 80's: The Critical Linkage of Personnel and Organizational Development Technologies"; "Computer Assisted Testing in Government Jobs"; "Implementing and Updating the Norfolk Police Department Performance Rating System"; and "The Administration of a Psychological Testing Program in Police and Correctional Agencies". Two invited addresses, three student papers and a paper by an invited speaker are included. (SLD)

ED337480

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Marianne Ernesto

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IPMA Assessment Council

PROCEEDINGS OF THE
1983 IPMAAC CONFERENCE
ON
PUBLIC PERSONNEL ASSESSMENT

MAY 22-26, 1983

WASHINGTON, D.C.

TM017294

Published and distributed by the International Personnel Management Association (IPMA).

Refer any questions to the Director of Assessment Services, IPMA,
1850 K Street, N.W., Suite 870, Washington, D.C. 20006 202/833-5860.

PROCEEDINGS OF THE 1983 IPMA ASSESSMENT COUNCIL CONFERENCE ON PUBLIC PERSONNEL MANAGEMENT

These PROCEEDINGS are published as a public service to encourage communication among assessment professionals about matters of mutual concern.

The PROCEEDINGS essentially summarize the presentations from information available to the Publications Committee of IPMAAC. Some presenters furnished papers which generally included extensions of their remarks, while others merely furnished a topical outline of their presentations. Tapes were also available for many sessions. Adequacy and detail of information available varied greatly. For just a few sessions no information was available from which a summary could be prepared.

Every attempt has been made to accurately represent each presentation. The PROCEEDINGS are summaries and condensations made by the reviewer(s). Persons wishing to quote results should consult directly with the author(s). In many cases extensive bibliographies were available which had to be excluded.

PREPARED UNDER THE GENERAL DIRECTION OF:

Clyde J. Lindley
Associate Director, Center for Psychological Service
Chair, Publications Committee, IPMAAC

Credit for major assistance in the compilation of the PROCEEDINGS goes to:

Thelma Hunt, George Washington University
Patricia Brennan, Part-time staff, Center for Psychological Service
Michael A. McDaniel, Montgomery County, Maryland
Mary Anne Nester, U.S. Office of Personnel Management

IPMA ASSESSMENT COUNCIL

The INTERNATIONAL PERSONNEL MANAGEMENT ASSOCIATION ASSESSMENT COUNCIL (IPMAAC) is a professional section of the International Personnel Management Association--United States for individuals actively engaged in or contributing to professional level public personnel assessment.

IPMAAC was formed in October 1976 to provide an organization that would fully meet the unique needs of public sector assessment professionals by:

- providing opportunities for professional development;
- defining appropriate assessment standards and methodology;
- increasing the involvement of assessment specialists in determining professional standards and practices;
- improving practices to assure equal employment opportunity and merit employment;
- assisting with the many legal challenges confronting assessment professionals; and
- coordinating assessment improvement efforts.

IPMAAC OBJECTIVES support the general objectives of the International Personnel Management Association--United States. IPMAAC encourages and gives direction to public personnel assessment; improves efforts in fields such as, but not limited to, selection, performance evaluation, training, and organization effectiveness; defines professional standards for public personnel assessment; and represents public policy relating to public personnel assessment practices.

IPMAAC EXECUTIVE COMMITTEE
Barbara A. Showers, President
Doris M. Maye, President-Elect
Charles B. Schultz, Past President

Published and distributed by the International Personnel Management Association Headquarters:

1850 K Street, N.W., Suite 870
Washington, D.C. 20006

(202) 833-5860

Refer any questions to Sandra Shoun, Director of Assessment Services.

DEDICATION

These proceedings of the Seventh Annual International Personnel Management Association Assessment Council Conference, May 1983, are dedicated to the memory of C. J. Bartlett, Robert L. Ebel, and Floyd L. Ruch. It is extremely tragic that three such important contributors to IPMAAC's origin, growth, and attaining of stature, should leave us within a six-month period.

C. J. (JACK) BARTLETT (1931-MAY 1983)

Throughout his career Jack Bartlett devoted much of his efforts to advancements related to IPMAAC's concerns - assessments in both the professional and scientific frames of reference. In his academic career he was influential in furthering the development of university programs for training personnel and psychometric psychologists. His Ph.D. students include many outstanding achievers in these fields.

ROBERT L. EBEL (1910-NOVEMBER 1982)

Dr. Ebel was a wholehearted supporter of IPMAAC. Several of his students have become officers and chairpersons in both IPMAAC as well as the regional consortia. In the early life of IPMAAC, he led significant workshops and was a featured speaker. He was a frequent resource to the regional consortia, offering both workshops and intensive training, including a videotape on test construction.

FLOYD L. RUCH (1903-NOVEMBER 1982)

Floyd L. Ruch was well known and respected in the field of industrial psychology as founder and President of Psychological Services, Inc.; as professor and head of the Business and Industrial Psychology Program at the University of Southern California; and as author of *Psychology and Life*, the second largest selling college textbook in *any* field published in this country.

Floyd Ruch will be long remembered by his many friends for his early pioneering work in the field of industrial psychology and for his professional contributions throughout his career. He was a frequent contributor to the Personnel Testing Council of Southern California.

The contributions of each of these individuals to the personnel profession will be remembered for many years to come.

TABLE OF CONTENTS

Page

PRESIDENTIAL ADDRESS - Is There a Future for Test Writers? 1

PAPER SESSION - Job Analysis Methodologies and Applications 5
Toward Multipurpose Job Analysis in a Large Public Agency: Rationale, Design and Progress 5
Use of CODAP to Develop a Selection Test Battery for Thirteen Federal Semiprofessional Occupations 7
Setting Minimum Qualifications: A Technology Driven Model for Determining the Need for Outside Expertise 8

KEYNOTE ADDRESS - Policies and Issues Facing EEOC Today 10

PAPER SESSION - An Introduction to Assessment Centers: Development and Uses 16
The Development and Usage of Assessment Centers 16
An Analysis of Common Assessment Center Dimensions 19

PAPER SESSION - Evaluating Other Minorities: The Handicapped and the Elderly 22
Development and Validation of a Manual Communication Examination for Selection Purposes 22
Performance Appraisal Systems and the Age Discrimination in Employment Act 23

SYMPOSIUM - Reasonable Accommodation in Selection and Employment of Deaf White-Collar Employees 26
Reasonable Accommodation in the Selection and Employment of Deaf White-Collar Workers 26
Reasonable Accommodation in the Selection and Employment of the Deaf 26
Testing Deaf Persons for Employment 27
Trends In Employment of Deaf White Collar Workers 29

PAPER SESSION - Behavioral Traits and Their Influence on Selection Procedures 30
Influence of Work Groups on the Selection Process 30
Providing Feedback to Individuals: Why It Doesn't Always Change Behavior 32
Leadership Effectiveness: A Management Perspective 33

PAPER SESSION - Assessment Centers: Research and Application 35
The Assessment Center As Psychological Process: An Analysis and Recommendations 35
An Examination of Internal Assessment Center Processes for Compliance with the Uniform Guidelines 38
Consultant-Agency Cooperation in Conducting Research On A Promotional Assessment Center for Police Lieutenant 39
Combining Multiple-Class and Class-Specific Assessment Center Exercises For Time and Cost Savings 42

Page

SYMPOSIUM - Development of a Valid Selection Procedure for Nineteen Professional Classes in a State Merit System 44
- Project Background 44
- Job Analysis 45
- The Search for an Appropriate Cognitive Abilities Test 47
- Test Development and Validation 49

SYMPOSIUM - Development of Job Related Valid Oral Board Examinations 52
- Development of Job Related Valid Oral Board Examinations 52
- Transfer Oral Board Procedures Used by the Marion County Sheriff's Department 55
- Oral Interviews as a Rating Tool for Librarians 56

SYMPOSIUM - Firefighter Selection: Non-Written Tests 59
- Physical Performance Tests: Setting the Pass Point 59
- Firefighter Physical Agility Examination City of Rochester - 1983 60
- Comments on Minneapolis Firefighter Exam 61

PAPER SESSION - The Effects of Recession, Labor Relations, and Reform of Public Agencies 63
- Labor Relations, Collective Bargaining, and Performance Appraisal in the Federal Government Under the Civil Service Reform Act of 1978 63
- Recessional Impact on Local Government Human Resource Management Systems: A Survey and Predictive Model 64

SYMPOSIUM - Designing and Implementing Personnel Functions in the 80's: The Critical Linkage of Personnel and Organizational Development Technologies 67
- The Critical Linkage of Personnel and Organizational Development Technologies 67
- Project Unity: Using OD Techniques for Implementation, Development and Capturing Employees Perceptions of a Performance Appraisal System 68
- The Use of Survey Based Approaches to Examine and Reduce Trainee Resistance to a Police Weapons Training Program 69

PAPER SESSION - New Developments in Selection for the Fire Service 71
- Validation of Fire Officer Promotion Procedures 71
- Problems in Job-Related Measurement of Reading Ability 71
- F. E. A. T. Revisited and Restructured 73

SYMPOSIUM - Computer Assisted Testing in Government Jobs 75
- Computer-Assisted Testing in Government and Industry 75

Page

PAPER SESSION - Biodata as a Public Sector Selection Strategy 76
Increasing the Odds for Producing Valid Biodata Instruments 76
Feasibility of a Content-Valid Biographical Questionnaire For the Selection of Municipal Police Officers 77

PAPER SESSION - Focus on the Economics of Testing: Measuring Utility 79
An Evaluation of Alternative Methods of Estimating the Standard Deviation of Job Performance to Determine the Utility of a Test In A Fixed-Treatment Sequential Employee Selection Process 79
Personnel Selection in the Wake of Teal: Proposals for the Defense of Noncompensatory Systems? 80
Test Utility for Mechanical Jobs 81
Comparison of Benefit Variance Estimation Procedures for Determining Utility 83

PAPER SESSION - Training and Experience Evaluations and Other Forms of Self-Assessment 85
Assessment for the Selection and Recruitment of Public Health Nurses 85
A Comparative Study of the Behavioral Consistency and Wholistic Judgment Methods of Job Applicant Training and Work Experience Evaluation 86
Reexamination of Self-Assessment for Employee Selection 88

PAPER SESSION - Exploring Issues Related to Fairness, Adverse Impact, and Test Bias 90
The Effect of Preselection on Adverse Impact Determination .. 90
Minimizing Adverse Impact While Maintaining a Merit System .. 91
Bias In Content-Valid Tests Revisited 92

PAPER SESSION - Assessing Managerial Skills 94
Development of Promotion Evaluation Procedures For A Centralized Referral System 94
Design and Implementation of the New York State Management Skills Inventory 95

PAPER SESSION - Alternatives and Traditional Selection Procedures 97
The Development and Evaluation of Three Traditional and Three Alternative (Non-Traditional) Selection Procedures 97
The Development of Some Innovative Predictors for Entry-Level Selection 100

PAPER SESSION - Measuring Productivity 104
Development of a Programming Productivity Measurement System........ 104
Development and Use of a Computer-Assisted (SAS) Goal-Setting System: Implications for White Collar Productivity 106

Page

INVITED ADDRESS - Alternative Uses of Traditional Selection Procedures 108

WINNER: STUDENT PAPER COMPETITION - A Statistical Analysis of Internal Sex Bias In A Job Evaluation Instrument 111

PRESENTATIONS BY STUDENT PAPER COMPETITION SEMI-FINALISTS -
- An Assessment of a Measure of Job Satisfaction in the Philippines and the United States 113
- Individuality in the Organizational Context: Conceptual Approach and Its Applications 114

INVITED ADDRESS - Recent Developments in Employment Law 116

INVITED SPEAKER - The Impact of Selection on Workforce Productivity and Output 119

PAPER SESSION - Physical Ability Testing 121
- Developing and Validation of A Physical Performance Test For Screening Vermont State Police Applicants 121
- Police Physical Ability Tests: Can They Ever Be Valid? 122

PAPER SESSION - The Development and Validation of Selection Standards for Law Enforcement 125
- Validity Generalization Results for Law Enforcement Occupations 125
- A Case of Increased Selection Test Validity By Assignment Of Job Relevant Language Points 126
- Rating Systems Are Not All Created Equal: The Baltimore Police Sergeant Experience 127

SYMPOSIUM - Implementing and Updating the Norfolk Police Department Performance Rating System 131
- Perceptions of the NPD Rating System 131
- Effects of Rater Training on Feedback and Goalsetting Behavior and Rating Errors: A Longitudinal Field Study 133

SYMPOSIUM - The Administration of a Psychological Testing Program in Police and Correctional Agencies 136
- Role of Inwald Personality Inventory and Minnesota Multiphasic Personality Inventory as Predictors of Correction Officers Job Performance by Race 136
- A Validation and Cross-Validation Study of Correction Officer Job Performance as Predicted by the IPI and MMPI ... 137

PRESIDENTIAL ADDRESS

Is There a Future for Test Writers?

Charles B. Schultz, Department of Personnel, State of Washington

There is a future for test writers.

Selection specialists more and more are getting into a variety of work related to selection in addition to writing multiple-choice tests. We are exploring various alternatives: improving ways to assess training and experience, improving ways to look at biodata, using the computer more in testing, doing a wide variety of research, and innovating. But my question is, specifically, "Are we going to be needed to write multiple-choice tests?" And my answer to that is"yes." I think multiple-choice tests need improvement. They need improvement to make them more valid. They need improvement so they will become more acceptable to the candidates and to the agencies they serve. They need to be made free from unwanted variance.

We have been hearing a great deal about utility research and validity generalization. No research program over the last few years has been more enlightening than the work of Frank Schmidt, John Hunter, and their various colleagues. We have much to learn from their research. But there is danger in interpreting it without enough sensitivity.

Utility research has shown us that old tests have high validity. Tests that have been around as long as I have seem to have about the same validities as tests that people are writing today. The work of Hunter and Schmidt has shown that you can take four ability tests and come up with about as high a multiple correlation as you can any other way. With a battery of four well known ability tests, you can get good, high validity coefficients for just about any job. So if we can predict validity with tests that hvae been around 50 years, why should I say we need to develop better tests?

For one thing, each time we can increase a validity coefficient by .01, we can increase productivity by $50 to $100 a year per candidate selected. It is my opinion that validities can be improved. We need better criteria against which to validate tests. We need to make tests more job related. We need to evaluate assumptions and alternative explanations in validity studies. We need to improve the wording of questions and instructions.

When you set out to do research on criterion-related validity, what is the biggest problem that you find? There are a number of big problems. There are the problems of restriction in range, of criterion unreliability, of sample size; but the biggest problem, in my opinion, is coming up with a good criterion measure. What Schmidt and his colleagues have shown is that the general ability tests do measure the criteria that have been used. And they have studied every criterion-related validity report they could get their hands on. But just what is the criterion that they are predicting?

The big thing I worry about when I'm writing a test is what I call unwanted variance. There are a number of kinds of unwanted variance: variance due to cultural bias, test-taking ability, particular experiences, and so on. But I'm talking about reliable variance that is not correlated with the ultimate criterion (distinguishing between the ultimate criterion and the criterion measure). The ultimate criterion is what you really want to measure; whether people are in fact doing a good job.

Many of you have done criterion-related validity studies. How many of you have been perfectly happy with your criterion measures in more than half of those cases? General ability tests that we have predict those criteria. Is that really what we want to predict? If you have something irrelevant in the criterion measure that is correlated with something irrelevant in your test, you may have a nice high validity. This is the familiar criterion contamination problem. I think that we need to do better criterion-related validity studies before we can rest on our laurels. In addition to the criterion problem, we have problems of assumptions, statistical corrections and interpretation.

In computing statistics, we make assumptions. Usually the assumptions don't really fit. The reason we make assumptions is that we can't compute statistics on what we have. We don't have enough information, so we make reasonable assumptions. We know there are various sources of error in our data, so we make corrections - based on assumptions. I have been wary about statistical corrections since the first time I made one on real data. I was merely correcting a single coefficient for unreliability. When I applied the correction formula, I came out with a correlation coefficient of 1.09. Later in my education, I applied analysis of covariance to groups that were unmatched on the independent variable. Of course, it undercorrected. Strange things happen when a correction is used. Corrections often compound unreliability.

Validity generalization studies attempt to prove the null hypothesis. Rather, they try to show that the size of the validity coefficient doesn't differ significantly from a prior hypothesis of the size. In short, the difference in validity from one situation to another is small enough that it could easily have occurred by chance. But the fact that the difference is not large compared to its standard error doesn't necessarily mean it's not real. Failing to reject the null hypothesis is not identical with accepting it.

I asked a number of colleagues and checked a number of sources to find out whether or not tests that are designed specifically for one job have higher validities than tests of general ability. The data are mixed on this. Many of you can point to individual studies where you find that the job specific test does have a higher validity. In these cases, the significant differences could be due to factors such as sampling error. Shall we conclude that they are?

If you have two tests that have equal validity, which one would you use--a general test or a specific test? I have some reasons for preferring the specific test. Do you have a friend who does well as a personnel specialist but would not do so well as an engineer, a librarian, an administrator, or a computer systems analyst? People generally do not do equally well at a number of different jobs. With job specific tests we can utilize individuals' specific talent more effectively. Research shows you can save millions of dollars in productivity by selecting on the basis of high scores on a general test. With more specific tests, you theoretically can save more millions and place workers into jobs that are a better fit.

Back in the '50s, Brogden showed the advantage of several test scores over a single test score when selecting a finite group of people for different jobs. As an illustration, let's say you have 20 highway engineer positions and 20 candidates to place who took an engineering test. What is the utility of using that test for these people? This is the limiting case. There are only 20 people so the selection ratio is 100%. You are going to have zero utility no matter how valid this general engineering test is. Suppose those 20 positions cover four specialties: design engineer, construction engineer, bridge engineer, and maintenance engineer. Suppose you could break that general engineering test down into subtests for each specialty. If these tests ranked the people differently, you could maximize the average rank on the specific subtests for the engineers assigned to each specialty. You can gain utility even if the validity of your individual subtests is low. You will usually not be placing the entire labor market. But if you choose your employees with a job specific test, it will leave persons with high general ability to fill other jobs for which they are particularly suited.

I advocate, for the future of test writers, both conducting more research to find what is going on and building better tests. What about our valid tests that have been around for 50 years? I wonder if they have problems. I wonder if some of you innovative test writers out there can do something with those tests to improve them. We get rid of any biases we can see, we reduce the reading level, we put tests into terms that the candidate will understand, we get rid of items that measure book knowledge but are not necessarily related to job performance. I believe these are more than cosmetic improvements.

Most of us working in public jurisdictions see some of these problems; we get some hunches; we have some intuitions. And we would like to collect some data. Most often our numbers are too small, our time is limited, and there is always the criterion problem. But if we are going to use job specific tests, we really need to improve them. I suspect that most of you from time to time do some test exchanging. Let's say that you send away to some of your colleagues for some tests. Let's say you get back six tests. When you look at these, do you say, "Oh, boy, I have 600 good questions I can use." Is that the typical response? Or is it more likely that, of those 600 questions, you get three questions you can use? And in those six tests there are 234 repeated items -- that have been around for 50 years.

In many tests that are used in public jurisdictions, the quality of English expression is an embarrassment. I talked with Jennifer French who is responsible for getting test questions into the Western Regional Item Bank. She gets questions from many jurisdictions, and she says there is a real problem with the editorial quality of those questions. When you have test items that are clumsy, empirical research may show those items seem to work pretty well anyway. That is, they are correlated with whatever our criterion is measuring. What effect do these test items have on the people who see them? What do managers in the agencies that you serve say? "Multiple-choice tests are okay at the entry level, but I wouldn't subject a professional to them. Why is that? Sometimes the tests we present to people look more like puzzles than like the assessments of the qualities we are trying to measure.

We need to do more thinking about the implications of research. Should we look at the results of research uncritically? Many times at this year's conference and last year's, I heard people say that we have gone about as far as we can go with multiple-choice testing, so we have to get into personality variables, lifestyle variables, and experience variables. They say that we have to come up with other measures in order to improve our predictions and our utility. I wholeheartedly agree that we should develop these other kinds of measures and should look at alternative selection devices. I also believe that we should spend a lot of time learning to become better test writers. We need to produce tests that are not an embarrassment; that are acceptable to our applicants and our agencies. Finally, rather than testing persons in the abstract, we need to write tests that measure differentially for different jobs.

PAPER SESSION

Job Analysis Methodologies and Applications

Chair: Wyverne L. Flatt, Directorate of Civilian Personnel, USAF
Discussant: Charles F. Sproule, Pennsylvania State Civil Service Commission

Toward Multipurpose Job Analysis in a Large Public Agency: Rationale, Design and Progress

Gary B. Brumback and Priscilla M. Palomino, U.S. Department of Health and Human Services

Rationale. The U.S. Department of Health and Human Services, a public agency with over 140,000 employees located throughout the country, has never had a job analysis methodology to endorse for use by the Department's component divisions. The Department has recently launched a major project to produce a cost effective job analysis methodology which can be applied at one time for multiple purposes and to produce model personnel procedures for the occupation on which the new methodology is pilot tested. The model procedures include position descriptions, classification guide, crediting plans for rating applicants, job elements and performance standards, and training criteria.

Design, Project Strategy and Organization. Because of a tight budget, we had to choose an in-house, boot-straps operation over contracting out the work. Although we have worried about in-house capabilities and availabilities for the project, not being able to thrust the work into the lap of a contractor may turn out to be a blessing in disguise. By developing the methodology slowly and painfully ourselves, learning and modifying as we go along, cajoling skeptics to come along, we may very well end up with a process where there is greater commitment to it because of greater involvement in its development.

In addition to the co-authors, the people in the network represent all the Department's components and 10 regions in the country. We have also established a Department-wide advisory group of some 25 subject matter experts (SMEs) in the pilot test occupation. An important function of this group will be to review the occupational content of our model personnel procedures.

Pilot Test Occupation. In order to pilot test the new job analysis methodology, we decided to conduct the pilot test in just one occupation so that we could best fulfill the second project objective of developing job-related personnel procedures.

The Program Analyst occupation emerged as the highest priority choice out of 400 possible occupations. Employees in this occupation are supposed to be doing just what the title of the series indicates, analyzing programs. There are about 2000 program analysts employed throughout the Department and in a range of grade levels. A significant reason for choosing the

Program Analyst occupation was that OPM has announced it soon intends to issue draft standards for that occupation. Our own job analysis of that occupation via the pilot test will give us much more agency-specific data than OPM obtained in its government-wide fact finding. Our own findings will thus help us determine how well OPM's draft standards represent our Program Analyst positions.

Pilot Test Methodology. The center of our approach is a multipurpose job analysis prototype. The core of the prototype is a job analysis procedure developed by the second author for the purpose of developing crediting plans, also known as rating schedules, for rating the education, training and experience of job applicants.

We have also drafted a basic position classification questionnaire. It contains operational definitions of the nine factors in the Federal government's Factor Evaluation System (FES) used for classifying positions and determining their grade levels. We plan to have the SMEs rate the target positions on the nine factors and to justify their ratings by citing job facts such as duties, tasks, KSAs, etc. determined in earlier steps of the prototype. We are also considering arraying the different pieces of already available job information, such as position descriptions, organizational charts and mission statements, work samples, etc. which classifiers typically examine in the course of their traditional desk audit approach. We may then try to make explicit the implicit linkages that are made between relevant pieces of information and factor level judgments based on them.

Besides the prototype, we will also be pilot testing a supplemental procedure. We have constructed and will shortly be administering to a random, cross-section sample of program analysts a standardized task inventory. We have an agreement with the U. S. Air Force for them to analyze the inventory data using their CODAP (comprehensive occupational data analysis program) which they pioneered and have been perfecting over the last few decades. We plan to gear our model personnel procedures toward the task profiles of the position families identified through CODAP's computerized clustering program.

Another supplemental procedure we are considering pilot testing is the Professional and Managerial Position Questionnaire (PMPQ) developed by Lt. Colonel Jimmy Mitchell of the Randolph Air Force Base. The PMPQ does have office face validity and contains several items which relate to some of the factors in the FES.

Progress. As you can tell, we are struggling along. When we wrote our proposal to talk to you today, we had thought we would be further along. Progress has been slow but we are determined to press ahead.

Use of CODAP to Develop a Selection Test Battery for Thirteen Federal Semiprofessional Occupations

Marvin H. Trattner, U.S. Office of Personnel Management

CODAP Description. CODAP is an occupational inventory analysis system developed at the U.S. Air Force Human Resources Laboratory. Typically, the CODAP inventory contains a listing of occupational tasks grouped into more general duty categories and also a background data section. Respondents provide demographic, education, training, and experience data in the background section. In the task section, respondents indicate which of the tasks in an occupation they perform and also the relative time spent in performance. They may also indicate the relative task importance and difficulty and whether the task has to be performed immediately upon entry into the occupation.

The CODAP system converts the relative time spent ratings into a proportion of total time spent score. Then individual and group job descriptions are produced. Also the overlap between all pairs of respondents in proportion of time spent in common task performance is calculated. The respondent overlap matrix is then cluster analyzed in order to uncover occupational subspecialties.

Development of Multioccupational Selection Battery. In order to perform the job analysis to develop a multioccupational selection battery, it was necessary to modify the CODAP procedure. Only the more general duties performed in the 13 occupations were listed. The duty descriptions emphasized the mental operations performed and occupational unique information was omitted if it had little bearing on the operations performed. In this way a single duty could apply to many occupations. The inventory was administered to incumbents in 13 Federal semiprofessional occupations.

Results. A representative sample consisting of 3.6% of the total Federal full-time population in the thirteen occupations responded to the inventory. Where the sample unexpectedly differed from the population it was probably because employees in the highest grade levels in the occupations were deliberately excluded. Consequently, the sample contained a greater proportion of minorities and women than the population.

The average overlap in proportion time spent among employees in different occupations was approximately .20; among employees in the same occupation it was approximately .30. The most time consuming duties for the total group when the occupations were equally weighted were clerical and administrative duties.

A cluster analysis was performed on the average overlap values among the occupations. One cluster contained the occupations concerned with processing of documents in connection with entitlement programs - contact representative, legal instruments examiner, general claims examiner, and loss and damage claims examiner. The other cluster contained occupations which provide technical support to key Federal occupations - management assistant, legal technician, procurement technician, and education and training technician.

The remaining step to be accomplished in assembling a selection battery for the class of occupations is to ask subject matter experts to rate the usefulness of specific knowledges, skills, abilities, and other characteristics for predicting successful performance of the most time consuming duties in each occupation.

The overlap between physically handicapped and nonhandicapped employees was calculated separately by occupation. The overlap analysis revealed that the physically handicapped resemble the nonhandicapped more than the nonhandicapped resemble each other.

CODAP can be used to test the consistency of the grade assignment system at an installation. It has been shown that within an occupation relative difficulty weighted by proportion time spent in task performance plus number of tasks performed is highly correlated with grade level.

Setting Minimum Qualifications: A Technology Driven Model for Determining the Need for Outside Expertise

R. Eugene Hughes, College of Business, West Virginia University

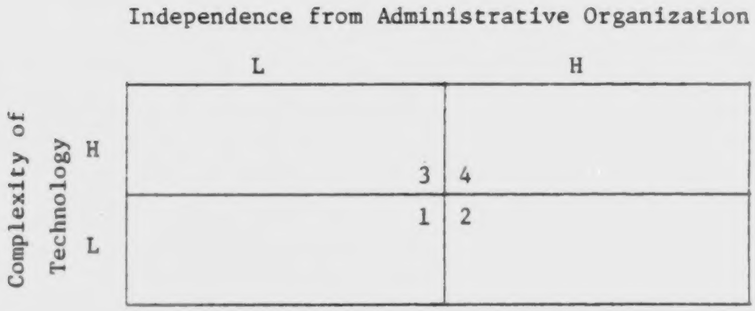
It is the responsibility of the organization's personnel unit to establish and control the job specification setting process so as to prevent the injection of nonperformance criteria into the job specification. It is quite possible, however, that when unique or complex technology jobs are being analyzed, the personnel unit may become unhealthily dependent upon the unit in which the job specifications are to be set. This dependence grows out of the personnel unit's inability to adequately evaluate job information developed and provided by the task unit that encompasses the job being analyzed. Since this dependence can result in some reduction of the personnel unit's ability to control the process, it is at this point that extreme care must be exercised in order to prevent the injection of nonperformance criteria. It is the purpose of the present paper to present a model that can be of value in describing situations in which this loss of control might occur.

Research shows that task units characterized as utilizing a complex technology are more effective if managed informally with low levels of centralization. There is also reason to believe that the more complex the technology, the more autonomous the task unit becomes in relation to the administrative function of the organization. The increased levels of autonomy can be expected to result in the emergence of a collegial form of self-management together with a set of norms and values unique to each such task unit. These unique sets of norms and values, when combined with a personnel unit that has an insufficient level of expertise in the task unit's technology tends to evolve into a threat to the objective determination of job specifications. In essence, the opportunity is created for the injection of nonperformance criteria that tend to serve only as mechanisms for the reinforcement of the task unit's norms and values.

A proposed model was presented so that the personnel unit can evaluate the circumstance surrounding the job analysis process so as to identify situations in which loss of control of the process is most likely to occur. (See Figure 1).

FIGURE 1

PROPOSED MODEL



Cell	Descriptions
1	Few problems - internal expertise is sufficient.
2	Some potential problems associate with norms and values. May suggest the need for a realignment of organizational authority.
3	Potential problems are technology related. Will require an objective inside expert or an outside consultant.
4	Significant potential for independent and combined problems that are based in the norms and values and technology of the work unit. Along with the response described for Cell 3, the expertise of a group-process consultant (inside or outside) will be required.

KEYNOTE ADDRESS

Policies and Issues Facing EEOC Today

Chair: Karen Coffee, Cooperative Personnel Services,
California State Personnel Board

Address by: Alvin Golub, Director, Office of Program Research
Equal Employment Opportunity Commission

Today, under the shadow of a soaring unemployment rate, the focal point of employee relations, more than ever, centers squarely on the selection process. Over the years, that process has been subjected to considerable challenge and refinement by a body of expertise directed to the achievement of rational employment and selection systems. Much of that focus sought to assure that applicant's abilities for jobs or promotions were geared to meeting business needs. No issue has received more judicial attention under Title VII of the CRA and the Age Discrimination in Employment Act than those which address hiring procedures, processes, criteria, qualifications and tests which have long been utilized for this purpose. As you know, the attention they receive by EEOC arises because of the resulting exclusion of blacks, Hispanics, women and older workers from certain jobs and their assignment to lesser positions. Yet these exclusions may, at times, be justified where there is an absence of the skills legitimately required by employers. Often, however, the exclusion is not justified but comes as the result of unfair and unlawful barriers which screen out certain candidates. These various selection devices have become the subject of the federal government's Uniform Guidelines on Employee Selection Procedures, probably the single most important document to emerge in the field of employment discrimination law. And it is not surprising that because of their widespread applicability, certain features of these Guidelines are the subject of considerable controversy. That controversy has surfaced on several fronts.

Some critics focus on the Commission itself. They contend, quite erroneously, that in its pursuit to obtain what is called "equality of results," that "EEOC has called for a virtual end to tests as conditions of employment."

But the assertion is false. The EEOC has not called for an end to tests or any other rational selection devices. Those assertions are erroneous or misleading in at least two significant respects. First, they confuse the role of the Commission with that of the courts. And second, they totally misconceive the nature and import of relevant case law with respect to selection procedures and the Federal Government's Uniform Guidelines on Employee Selection Procedures.

As to the role of the Commission, it must be emphasized that EEOC has no authority, statutory or otherwise, to require employers to undertake or discontinue practices. Under the statute, only the courts have such authority. And it is the case law resulting from judicial decisions that has, for the most part, defined the parameters of Title VII and the Age Act with respect to employment policies and such selection devices as tests. The Supreme Court in the Griggs decision took an early lead in requiring that employment selection procedures which operate to exclude disproportionate numbers of minorities and/or women from job opportunities must be job-related and must be predicated upon business necessity in order to justify their continued use. The language of the court strikes home again in McDonnell vs. Green where it refers to:

> "standardized testing devices,...neutral on their face, (operating) to exclude many blacks who were capable of performing effectively in the desired positions. Griggs was rightly concerned that childhood deficiencies in the education and background of minority citizens, resulting from forces beyond their control, not be allowed to work a cumulative and invidious burden on such citizens for the remainder of their lives."

These Supreme Court cases serve as the foundation for Commission guidelines along with many circuit court decisions involving employment selection procedures. And the Congress has consistently endorsed the concepts embodied in these cases. The legislative history of the 1972 Amendments to Title VII, for example, is replete with references to the Griggs decision. The debates and committee reports bear witness to the fact that Congress intended that employment practices having an adverse impact on the career opportunities of women and minorities, thereby operating to freeze in the status quo of discrimination, should be subjected to close scrutiny and utilized only where they serve a job-related business necessity. With that underpinning, we believe that EEOC and the other agencies acted fairly and in concert with the intent of Congress in seeking to eliminate discrimination which results from the use of arbitrary and non-job-related employment selection devices.

The further assertion that the Commission has called for a "virtual end to tests is, we believe, also an incorrect reading of the substance of the Uniform Guidelines on Employee Selection Procedures. These Guidelines simply describe the circumstances under which tests and other policies would be considered to have an adverse effect on the employment prospects of women and/or minorities and the standards which the federal enforcement agencies will use to determine whether such procedures have been demonstrated to be job-related. The Guidelines are a far cry from a requirement that employers discontinue the use of tests as selection devices. Indeed, once a procedure is shown to be job-related, it may be used even though it may have an adverse impact on the job opportunities of women and/or minorities, unless a substantially equally valid procedure with lesser adverse impact is available to the employer or unless the procedure is intentionally used to discriminate on the basis of race, sex or ethnic group membership.

The job-relatedness requirement in the selection procedures is dealt with squarely in the Griggs decision:

> "(Title VII) proscribes not only overt discrimination but also practices that are fair in form, but discriminatory in operation. The touchstone is business necessity. If an employment practice which operates to exclude (minorities or women) cannot be shown to be related to job performance, the practice is prohibited."

The Guidelines also contain a "bottom line" provision -- a failsafe mechanism -- that underscores the government's emphasis on encouraging good faith efforts. This provision advises employers and others that, even if their selection device discriminates unlawfully, the government will not take enforcement actions which require that the device be *validated* provided that the total selection process does not have an adverse impact. The "bottom line" of the testing guidelines represents another step by government to encourage voluntary efforts as a preferable alternative to costly court enforcement.

A more subtle and more sophisticated challenge to the *Uniform Guidelines* has been raised recently by some in the academic community. They do not suggest that EEOC has acted beyond its statutory mandate. Rather, their argument is that "all cognitive ability tests" are valid for all jobs, "regardless of job level or complexity." Thus, no test is invalid for any job. The proponents of this theory concede that degrees of validity may vary among test and that some tests may be better predictors of job performance than others. The essence of their argument, however, is that almost any test is better than *no* test, regardless of its impact in terms of excluding minorities, women and older workers from certain jobs.

Prompted in part by research supportive of the theory, but never apparently subjected to independent verification, the APA Committee on Psychological Tests and Assessment has in the past suggested to the Commission that it take "preliminary steps to revise the *Uniform Guidelines* to reflect the current body of empirical knowledge (and) recent theoretical developments." But such a move at that time on the part of the Commission was unwarranted. Indeed, only recently the same APA Committee commented on the *Uniform Guidelines*, that "these Guidelines have attained consistency with the Standards in those areas in which comparisons can be meaningfully made." Nor does the developing research enjoy the universal support of the psychological community. The February 1983 draft of the *Joint Technical Standards for Educational and Psychological Testing* states: "there is no basis in evidence for a universal a priori assumption that all cognitive ability tests are valid for all jobs, for all groups of people or for all situations." (p. 2-12). Furthermore, while the Guidelines commit the Commission to an evaluation of "new strategies for showing the validity of selection procedures...as they become accepted by the psychological community," the EEOC must also be cognizant of the Supreme Court's reminder in *General Electric Co*. v. *Gilbert*, 429 U.S. 125, 12EPD para.11, 240 (1977) that it will decline to "follow administrative guidelines...where they conflict with earlier pronouncements of the agency." 429 U.S. 143.

It seems important to repeat that through the years, the process of developing selection guidelines has been, we believe, a rational one, culminating ultimately in the 1978 adoption of the Uniform Guidelines by four enforcement agencies EEOC, OFCCP, the Department of Justice and the then Civil Service Commission. Court decisions laid the groundwork for the development of these Guidelines. Courts, in turn, have looked to them in determining the validity of selection devices. While refinements were appropriate on occasion, the EEOC has never deviated from its original position subsequently affirmed by Congress and the courts. Nor does the Commission feel that it should embark on costly and possibly unwarranted revisions to its position unless and until Congress or the courts or the clear weight of indisputable proven facts dictate such an overhaul.

An even more critical reason to reject the suggestion that recent theoretical developments about "validity generalization" be incorporated into the Guidelines at this time is that the upshot of such an action would be to encumber and obstruct our objective of enforcing workplace equality. The argument that any cognitive ability test, because it can be assumed to be valid, may be used with impunity notwithstanding its exclusionary effects, puts equal opportunity enforcement back on square one. It appears to be an argument that Title VII reaches only "intentional" discrimination. What it ignores is that employers are in exclusive control when it comes to most selection devices used to screen applicants for jobs. Courts have, understandably, put to employers the task of justifying the premises on which these selection devices are applied. Once it is demonstrated that in using a device, certain types of people are being screened out, that justification process must begin.

The level of scrutiny applied to each procedure may vary, depending on the extent of its use, the degree to which it is job-related, whether it has an exclusionary effect, and the level of control exercised. However, where exclusion occurs, no selection procedure should be presumed sacred. It is, perhaps, for this reason that courts have generally rejected the "validity generalization" argument.

Turning now to the debate over race and sex conscious affirmative action plans, the issue cannot be easily addressed by slick, overly simplistic explanations and solutions. There are few who dispute the fact that Congress has consistently endorsed the notion of affirmative action by federal contractors as a vital tool in opening job opportunities for women and minorities. Presidents, going back to the 1940's, found it appropriate to use the government's procurement authority in seeking to do the very same thing. And presidents, more recently, have seen fit to appoint a black and a woman to the Supreme Court because it was obvious that blacks and women had been previously excluded. Does anyone doubt that these were race and sex-conscious decisions? Were those Presidents wrong?

The Supreme Court, in the Weber case, bserved that the Kaiser Company plan was "an affirmative action plan voluntarily adopted by private parties to eliminate traditional patterns of racial segregation."...."The very statutory words were intended as a spur or catalyst to cause employers and unions to self-examine and to self-evaluate their employment practices and to... eliminate, as far as possible, the last vestiges... of discrimination." The court goes on to say that, "We, therefore, hold that Title VII's prohibition against racial discrimination (See 703(a) & (d)) does not condemn all private, voluntary race-conscious affirmative action plans." In other words, why wait for the government to intrude, to investigate, to make a finding of discrimination before taking necessary action to change exclusionary patterns and practices?

The Court went further by saying that employers who act reasonably, not only to correct deficiencies but to create opportunities for those historically excluded, need have no fear that they will be held liable for claims of reverse discrimination. Employers and unions are, it seems, free to institute such plans on their own without government prodding, government interference or government tribunals.

Recent data from the Census bureau and EEOC's own surveys describe gains by minorities and women in the 1970's into occupations from which they had been largely excluded, although representation in high-paying jobs was still low. Still, a record number of women became lawyers, doctors, ministers, architects, security guards, construction workers and computer operators. Gains were also made by blacks and Hispanics.

Among other reasons, all this occurred because many business and union officials recognized the need to open the American workplace to those who were previously excluded. Race and sex-conscious decisions were made to right a wrong and we were all the beneficiaries of those decisions. What's more, there was little challenge raised on the basis of reverse discrimination because affirmative action became part of a national way of doing business and employers, employees and unions knew it was the right thing to do.

In the final analysis, the real objection to affirmative action seems to land squarely on the concept itself. Admittedly, the objections cannot and should not be dismissed easily. Where plans or actions, under the guise of affirmative action, violate Title VII by discriminating, reverse or otherwise -- there is no immunity from the full enforcement of the law.

Earlier this year the Commission voted to submit an *amicus curiae* brief in an affirmative action case involving a municipal employer. The case I am referring to is *Williams* v. *City of New Orleans* (5th Cir. No. 82-425). In *Williams*, the Department of Justice had filed a brief which argued that affirmative relief violated Title VII as well as the constitution. The Commission intended to submit a brief supporting affirmative action because, among other things, judicial ratification of the Justice Department's position would undermine the Commission's Guidelines, settlements, consent decrees and court orders providing for affirmative relief. It would prevent employers and other entities from using non-controversial recruitment and training efforts as well as flexible numerical goals and timetables. The _Williams_ case became, and still is a cause celebre. What is ignored in all the controversy is the regular and every day application of affirmative action. It has been used in hundreds of cases that have come before the Commission. About 450 out of 2,000 EEOC cases which were settled or tried between 1977 and 1982 resulted in settlements or court orders which provided for affirmative relief. Currently, the Commission is also monitoring consent decrees which provided for affirmative relief in 20 cases alleging "systemic," or patterns or practices of broad discrimination against minorities and/or women.

In addition, the Commission submitted briefs in approximately 26 Federal and state cases between 1972 and 1982 which supported the legality of affirmative action in general, or the legality of specific affirmative action plans. In 10 of the 26 cases, the Commission appeared as a party. In 16 of the 26 cases, the Commission appeared as _amicus curiae_. Of the 16 _amicus_ cases, 10 involved state or municipal employees. Yet there was little publicity, media coverage or suits filed challenging affirmative action efforts. The debate will, no doubt, linger on.

Finally, I am aware that the subject of employment discrimination has been discussed over and over again and I have no illusions that there is any single final resolution. But there are solutions which make sense and which offer a significantly better society and world than we have today. I, personally, have little hope for such a world unless the persistent issue of employment discrimination and economic inequality is dealt with by all of us.

PAPER SESSION

An Introduction to Assessment Centers: Development and Uses

Chair: Dennis A. Joiner, Dennis A. Joiner Associates
Discussant: Anita D. Collins, State of Florida

The Development and Usage of Assessment Centers

Carol Granfield, City of Fairfax, Virginia

An assessment center is a validated and observable process which is developed for the purpose of evaluating supervisory and management ability. It identifies strengths and weaknesses of its participants through an objective process. Assessment centers consist of a series of exercises specifically designed to replicate critical elements that are necessary for a particular job. Although the exercises are not identical to the job, they do involve job related tasks, skills and abilities used in the target position.

Participants are observed by several trained observers/assessors in a controlled environment. An administrator oversees the entire program and directs both participants and assessors. A variety of individual and group exercises are included in the composition of an assessment center.

Assessment centers are being used more and more by both the private and public sector. The federal government has utilized various types of assessment centers for several years. At the local level a primary usage of recent years has been that of fire and police promotional processes. This particular usage will be discussed in great depth later in this paper and will detail its experience in the City of Fairfax, Virginia, as well as pilot testing that was accomplished in other localities.

Assessment centers may be used at entry level, for career development or for management and promotional purposes. They are becoming more utilized for the selection of key managers such as at the local level as an aid in the selection of a city or county manager or a department head.

Development of an Assessment Center. The first and primary step in developing an assessment center is to conduct a comprehensive job analysis of the target position in order to determine the dimensions that are critical for success.

The system that was utilized in the City of Fairfax was that of initially having all supervisory employees complete a detailed questionnaire which required the itemization of tasks performed listing priorities of tasks, their duration, frequency and importance. Additionally, a section required the listing of knowledge, skills and abilities (KSAs) necessary to perform the duties of the position. The next step was to identify the dimensions or critical elements of the job related to career development or promotion.

The next phase in the development of an assessment center is the development of exercises. There are essentially two types of exercises that are utilized in assessment centers which are individual and group exercises.

In the category of individual exercises a primary exercise that is usually included is the "In-Basket." This exercise simulates the paperwork, with the level of difficulty appropriate, that the target position would be exposed to during the normal course of duties.

Another individual exercise frequently used is an "Oral Presentation". This exercise generally consists of a stand up oral presentation which is based upon a problem, task, or topic given to each participant prior to the exercise.

The group exercises are usually called "Leaderless Group Discussions." This type of exercise involves the participation of a number of candidates in a group discussion and allows assessors to observe and evaluate a candidates's behavior in a group situation. In most cases the group is a task force assigned to solve a problem. Another type of group exercise is one which involves role-playing. In this type of exercise each candidate is given a specific role to play within the group. Again this group is observed by multiple assessors.

Both the individual and group exercises in the case for promotional purposes are researched, studied and evaluated with the use of experts in the field, i.e., police or fire personnel knowledgeable in the particular field. Thus, in the case of assessment centers for public safety promotional purposes, the content validity of the process is easily seen by the participants and developers and therefore is acceptable to most.

Training of Assessors. The next key concern is the training of the assessors as it is an integral part of a successful assessment center. The training segment is highly important in the case of promotional assessment centers; the assessors play an important part in the careers of many people.

There are four basic steps in which assessors are trained: observe, record, connect with traits, and rate. The assessors are provided a notebook that includes all of the exercises, evaluation techniques, schedules, and guidelines to be followed. They are trained to evaluate two or more participants at one time during an exercise and they are trained on how to handle a variety of situations that may arise based on past experiences. A rating scale of 1 to 9 is used and assessors are trained to objectively observe the candidates, record what is being said and done to connect this to the dimensions being evaluated and, finally, to relate the 1 to 9 range of rating.

An assessment center that is utilized as part of a promotion process will have each dimension receive a score as well as have an overall score based on all dimensions which will be used as the final score for promotion. Based on a comprehensive job analysis, varying weights of importance are given to each dimension that is being evaluated. A higher weight is given to a dimension that is of critical importance to the position than to a dimension that is necessary, but not critical. The final score is converted to one on a scale of 100. Fairfax has not utilized a pass or fail score because the assessment center is only one portion of the public safety promotion process. Studies do indicate that a score of 70 or better could be the passing point if one were used.

Feedback. The feedback that is provided to the participant may be in a report form and/or include a session with the administrator or designee to go over the management strengths and weaknesses of the participant.

In a career development assessment center, the feedback session is the most important part of the assessment because it provides the participant with what is necessary to further develop his or her career. The feedback provided allows the participants to better understand themselves as well as to better prepare themselves for upcoming endeavors in the management field.

Fairfax Experience. Based on an analysis of the initial assessment center experience in Fairfax, some adjustments have been made over the years. Much of it, however, remains the same. The job analysis of target positions is conducted with dimensions necessary identified. Additionally, the assessor training is quite comprehensive.

The assessment centers conducted for both fire and police are generally one-half day in length for most ranks. The exercises consist primarily of two individual exercises (an in-basket and an oral presentation) and two leaderless group exercises. All exercises are developed to be either police or fire-related as applicable.

Assessment centers have been used as one portion of the promotion process within the fire and police departments. They have been given a value in weight of 40% of the overall process towards the final score. At the conclusion of the process, all participants are given a comprehensive feedback session. Based on participants pursuing recommendations made during these sessions, significant improvements have been noted in weak areas at subsequent assessment centers.

Since 1980, when Fairfax initiated the use of assessment centers for public safety promotions, the promotions have been made with no law suits in connection with the process. The assessment center offers public safety departments an effective promotional process. The key to the success of an assessment is the quality and training of the assessors, the quality of the exercises and evaluation technique and the feedback that is provided to the participants relative to their strengths and weaknesses.

Career Development. Assessment centers can also be utilized for training and career development purposes. The assessors for the career development center will soon be trained. All Fairfax department heads and key supervisory personnel are scheduled to be trained. Their benefit will be two-fold, as in addition to assisting as assessors of employees, they will learn to objectively evaluate employees. This tool will be most helpful in the area of performance evaluation.

Having administered, developed and participated in many assessment centers, I am convinced that the assessment center concept is viewed by employees and management as an excellent means to evaluate supervisory potential. Further, I believe that Fairfax has proven that it can be cost effective in small organizations if you have the in-house expertise, innovation, motivation and training capabilities.

An Analysis of Common Assessment Center Dimensions

Patrick T. Maher, Personnel and Organization Development
Consultants, Inc., LaPalma, California

The assessment center method requires that assessors evaluate behavior by placing it into assessment dimensions. Such dimensions, in turn are defined operationally. The initial emphasis was limited to analyzing the more commonly used dimensions in supervisory and management assessment centers. The purpose of this study, then, was limited to identifying as many different elements in use so that persons defining dimensions have a comprehensive resource. These dimensions are more commonly known in personnel testing as KSAs, KSAPs, or KSAOs, depending on one's preference.

Dimensions normally consist of two components: (1) the title or name of the dimension (e.g., Oral Communication Ability) and, (2) the operational definition (e.g., effectiveness of expression in individual or group situations).

In reviewing the various lists of dimensions used in different assessment centers, some apparent trends seemed to emerge. Titles were frequently the same, but the operational definitions were partially or entirely different. Definitions were similar, or even identical, but the titles were different. Some assessment centers merely used a dimension title without an operational definition. Each operational definition consisted of one or more elements. Identifying each element simplified the ability to compare the similiarities and differences of each dimension, regardless of title or definition.

Problems. First, many dimensions are loosely used in assessment centers That is, boilerplating of dimensions is probably the true source of some assessment center dimensions rather than a job analysis. Second, there is a danger that assessors not properly trained in the assessment center method, particularly in the area dealing with placing behaviors into dimensions and rating dimensions based on listed behavior, will improperly record behavior and rate dimensions.

Another danger is in the training of assessors. There appears to be a common myth that anyone who has been through an assessment center is a qualified expert and can than train others as assessors. Given the many differences in the various definitions for the various dimensions, it is unlikely that a person who has been an assessor once can become the ideal trainer for others. It will take considerable training and experience to fully appreciate the subtle, yet significant changes that may exist between various dimensions similarily titled and defined, and yet actually quite different.

Persons conducting job analyses, whether for the purpose of constructing assessment centers or for any other reason, will obtain a resource of common management or supervisory KSAs. By identifying each of the common elements, dimensions can be defined that meet the specific needs of the target job. Training of assessors in the meaning of dimensions, and enabling them to place behavior into the proper dimension can be simplified. Perhaps one of the hardest tasks for any assessor is determining which dimension to place a specific behavior. Our experience has been that by listing the various elements found in a given dimension, assessors have a better understanding of the relationship between a given behavior and a given dimension.

Leadership. The dimension of "leadership" provided some unique problems. It was the one dimension that had the most elements contained in it. More significantly, it is the one dimension that was most unique in its definition. Seldom did the various definitions repeat themselves or involve identical elements.

Leadership is discussed as being autocratic, democratic, dynamic, inspirational, and telepathic. It is viewed as passive and active. In reality, leadership involves the ability to communicate, to be independent, to make decisions, to plan and organize the work of one self and others, to analyze problems, to take risks, to be self-starting, flexible, and sensitive, and to have a high degree of stress tolerance.

Given these problems, leadership is best eliminated as a specific dimension, and, where it is used, should probably be treated as a construct.

Judgment. Judgment (also titled decision making) is generally regarded, at least by the courts, as a construct, and therefore not appropriately validated by content validation. Since most assessment centers use content validation as a validation strategy, it would appear to be difficult, if not impossible, to use judgment as a dimension of an assessment center. To still retain the definition of the traits commonly labeled as judgment some assessment centers have used "decision making" as the title, but that is probably not much better than judgment. Another method of still retaining the traits contained in the dimension might be to combine them with the traits found in problem analysis and work perspective. No matter which method is used, it would be best to avoid the titles "judgment" and "decision making."

Number of Dimensions. In reviewing the dimensions for actual assessment centers (N=28), there was also a wide range in the number of dimensions being evaluated. The number of dimensions ranged from 5 to 19, with the mode being 9 and the mean being 11, with 64 percent having between 9 and 14 dimensions being evaluated.

A subjective analysis of the dimensions indicate that where there are fewer definitions (less than 8), the definitions tended to be more complex; while the definitions tended to be simpler and more discrete where there were more than 13 dimensions.

Assessor Background. There is a common belief that assessors must share a common background with the position being assessed; that is, only chief officers of the fire service can properly assess a participant in the fire service, and so on. In reviewing the elements of common assessment center dimensions, this perspective does not seem to be supported. Assessment center dimensions measure generic management skills. Nothing in any of the dimensions require specific job or occupational knowledge or experience. Indeed, an assessor with a broad management background would probably have a better perspective than would an assessor who had management experience that was limited to a specialized field.

Summary. There is still need for extensive research into the nature of dimensions used in assessment centers. One area is the need to match definitions of common management and supervisory skills and abilities with the professional literature in the field.

There is also a need to determine the extent to which different definitions of the same general ability (e.g., planning and organizing) impact a participant's rating in an assessment center. If various definitions have no substantial impact on how a person is rated, then a standardized definition could be developed and used. If, however, different definitions result in different performance ratings for the same person under the same circumstances, then a properly developed definition becomes much more critical to the process.

The current research, however, does indicate that there are differences at the present, and until there is clear evidence to the contrary, each assessment center must be based on a carefully constructed job analysis.

PAPER SESSION

Evaluating Other Minorities: The Handicapped and the Elderly

Chair: David Lipsey, Training and Development Services

Development and Validation of a Manual Communication Examination for Selection Purposes

William E. Tomes and Nancy Whitlock, South Carolina State Personnel Division; and Maureen Irons, South Carolina School for the Deaf and Blind

As specialists in deaf education began to adopt a philosophy of Total Communication, which encourages maximal communication through any means, they began integrating sign language into programs for educating deaf students and instructional personnel.

Staff members of the South Carolina School for the Deaf and Blind have been using sign language to communicate with students for a number of years. However, some of the non-instructional staff who had been working at the school prior to introduction of sign language into the deaf education program had not been trained in its use. Realizing the need for all staff who communicate with the students to have sign language skills, the school implemented a program to make sign language training available to the staff.

The school staff had identified twelve job families of employees who, in their opinion, needed the highest level (level one) of sign language skills. Using information from position questionnaires, on-site observations and personal interviews with employees, we developed a task inventory and a list of special skills and abilities used in performing the tasks. The tasks were divided into six major duty areas.

A total of one hundred and seven employees of the Deaf and Blind School completed the questionnaire. Four of the job families indicated that they used the skills less than 20% of the time, while the remaining eight job families used the skills at least one-third (33%) of the time. In addition to importance and time ratings, we also identified the specific tasks in which manual communication skills are needed. Not surprisingly, most of these tasks are in the Student/Client Interaction duty area. An analysis of the data also showed that expressive skills are used more than receptive skills.

We recommended to the school staff that to ensure content validity, the assessment instruments developed should be based on the tasks in which the job analysis whoed manual communication skills are needed.

A survey of state institutions for deaf education showed that although none of the respondents require sign language skills as an absolute minimum requirement at entry, most do require the acquisition of sign language skills after a certain period on the job, ranging from six months to three years. Skills are assessed by various methods ranging from informal evaluations by supervisors to formal reviews by a panel of staff members. Tools used in the assessment also vary.

In 1979 the Manual Communication Committee of the South Carolina School for the Deaf and Blind, after extensive research, chose "Signed English" to be the official system of the school.

After three years of staff training, an expressive and receptive skills evaluation method was developed by the Manual Communication Committee. The Committee decided to have the individual sign a story presented on a transparency on an overhead projector. Each story contains target words or phrases designed to measure performance on specified objectives, such as correct use of signs of fingerspelling, speed, and facial expressions. The individual is videotaped and each videotape is viewed and scored on each objective by a team of two certified interpreters and one deaf individual.

Because the employees reported using about 60% expressive skills and 40% receptive skills when using sign language, the expressive part has a weight of 60% and the receptive a weight of 40%. Maximum raw scores are 30 on the expressive part and 20 on the receptive part of the signing exercise. The total raw scores on expressive and receptive skills are added together, converted to a percentage score, and then these percentage scores are assigned a corresponding proficiency level.

As employees become more skilled in the use of sign language, proficiency levels may need to be adjusted and alternate versions of the assessment instrument developed.

Performance Appraisal Systems and the Age Discrimination in Employment Act

Michael H. Schuster and Christopher S. Miller, Syracuse University

Congress enacted the Age Discrimination in Employment Act (1, Sec. 621-34) "to prohibit discrimination in employment on account of age in such matters as hiring, job retention, compensation, and other terms and conditions of employment" (2, p. 976). The legislation was targeted to promote the employment of older persons based upon their ability rather than their age, prohibit arbitrary age discrimination in employment, and assist employers and workers in finding ways of meeting the problems arising from the impact of age in employment.

The analysis on employer's formal personnel evaluation records will examine how older employees are evaluated, the use of an employer's appraisals and evaluations of the older employee's on-the-job performance and the roles such evaluation play in federal court decisions. Also assessed is whether the performance of older employees is being evaluated by the fairest and most up-to-date methods of performance appraisal available.

Performance appraisal is the systematic evaluation of a worker's job. The most common performance evaluation methods include: (1) graphic rating scales; (2) employee comparisons; (3) check-lists; (4) free form essays; and (5) critical incidents. Since many organizations do not use formal methods of evaluation, informal or ad hoc methods must also be considered. An analysis of each of the above methods was presented.

The fact that an employer has conducted some type of appraisal of an aggrieved employee's job performance becomes significant once the complainant in an ADEA suit has established a prima facie case of age discrimination, which is the minimum level of proof an ADEA plaintiff must offer to avoid having his/her claim dismissed and to shift the burden of evidence to the employer. Three elements are generally required to establish a prima facie case: (1) employee's membership in the protected group, (2) his/her discharge and (3) his/her ability to do the job (4, p. 1125-1124).

Evidence for a prima facie case may consist solely of specific incidents of discriminatory conduct, or there may be a combination of discriminatory conduct and statistical evidence. Once the plaintiff has made out a prima facie case, the defendant has the burden of going forward with evidence that reasonable factors, other than age, were the basis for the alleged discriminatory employment practices.

It has become increasingly clear that the use of an employer's formal personnel evaluation records can play a critical role in the decision making process of the federal courts. Formal records are not a requisite for rebutting the prima facie case. Such rebuttal can also be accomplished through the testimony of fellow workers and superiors. However, where there are no formal records to substantiate such testimony, the attorney for the older worker may attack the credibility of the employer's witnesses, thus discounting the value of the only source of appraisal-related information. This holds particular significance for a jury trial, where the jury may be somewhat sympathetic to the employee.

Even when formal performance evaluations exist, they will be of little evidentiary value unless the older employee has been appraised in terms of definite identifiable criteria based on the quality and quantity of his/her work. A good example of how a court will react to a well-developed system of evaluation is found in Stringfellow vs. Monsanto, where the district court judge was impressed that the employer had utilized techniques and criteria for performance evaluation published by the American Management Association. As will be seen, the judge in Stringfellow is not alone in giving substantial weight and high regard for an employer's genuine effort at individual job performance appraisal. In order to protect the jobs of older workers, it is this type of fair and reasonable performance evaluation envisioned by the proponents of the ADEA.

The issue of job performance has been critical in three types of ADEA actions: promotions, layoffs/retirements, and discharges. This paper addresses twenty-six ADEA cases decided in the federal courts where the evaluation of the employee's performance has been a determinative factor. Three of these decisions concern employees complaining that their failure to receive sought after promotions was due to their age and not their on-the-job performance. Eight decisions involved plaintiffs complaining that they were laid off or involuntarily retired based on their age.

The remaining fifteen decisions involved outright discharges. An analysis of these cases and court decisions was presented.

Conclusions. The type of personnel action appears to dictate the nature of the proof required to substantiate a nondiscriminatory employer decision. Promotion decisions require that the employer only show that the complaining employee was not as qualified as the candidate selected for an expanded role in the organization. Along the same lines, layoffs and retirements require the employee to demonstrate that the laid-off/retired employee was not as qualified as those selected to remain. In contrast, a discharge decision will not be upheld where the employee has performed at a minimally acceptable level. Therefore, discharge actions will probably require an expanded justification by the employer in order to establish that the decision was made on a nondiscriminatory basis.

Formal performance evaluation procedures have not been required for an employer-defendant to mount a successful defense. The courts have permitted less reliable sources of employee performance information to be used as conclusive evidence substantiating an employer claim of nondiscriminatory decision-making. However, an employer that conducts periodic well-designed performance appraisals and makes personnel decisions based upon the performance appraisal is likely to successfully rebut a claim of discriminatory conduct.

It is evident that the use of fair and consistent performance appraisal methods supports the intent of the ADEA to place older workers on an equal footing with their younger counterparts. It is not necessarily clear, though, that employees are being successful in demonstrating this proposition. Of the 26 ADEA cases analyzed, the presence of a well-structured, regularly conducted performance evaluation system was found just 34% of the time.

SYMPOSIUM

Reasonable Accommodation in Selection and Employment of Deaf White-Collar Employees

Chair: Dorothy J. Steffanic, U.S. Office of Personnel Management
Discussant: Gary Viall, Small Business Administration

Reasonable Accommodation in the Selection and Employment of Deaf White-Collar Workers

Anice V. Nelson, U.S. Office of Personnel Management

The Federal Government's concern with the employment of deaf individuals can be traced back to 1908, when President Theodore Roosevelt signed an Executive Order permitting the admission of deaf persons to civil service examinations. Since that time there has been a progression of programs and laws concerned with employment of handicapped persons. In the period up to the end of World War II, efforts were made to use job analysis techniques to identify jobs suitable for deaf and other physically handicapped persons. In 1948, concern about employer reluctance and attitudes towards persons with physical disabilities led Congress to amend the Civil Service Act to prohibit discrimination against handicapped persons in Federal employment. In the 1950's and 60's, employment of the deaf and other physically handicapped persons was facilitated through the use of modified testing procedures and special appointing authorities.

The 1970's saw many new initiatives in Federal employment of handicapped persons. Among the most important were the Rehabilitation Act of 1973, which required affirmative action in the employment of handicapped persons in Federal agencies, and the discrimination complaints procedure issued by the Civil Service Commission in 1978, which introduced the requirement for reasonable accommodation in Federal employment of handicapped persons. Of particular significance for the deaf was a provision of the Civil Service Reform Act of 1978, which provided Federal agencies with the authority to pay for interpreters for their deaf employees.

Reasonable Accommodation in the Selection and Employment of the Deaf

Dorothy J. Steffanic, U.S. Office of Personnel Management

The issue of reasonable accommodation for deaf people is quite complex in that it involves virtually every step of the selection and employment process, including employment tests, interviews, training, daily communication needs, and job duties. Reasonable accommodation is designed to alleviate the problems of language and communication which are inherent to deafness. Prelingual deafness (deafness which occurs before the individual acquires a spoken language) poses the greatest problem because it usually prevents the individual from developing real mastery of the language of the hearing community. Instead, the prelingually deaf person tends to develop fluency in a manual sign language, which is linguistically different from spoken languages.

The communication problems created by deafness can be ameliorated through a variety of measures. Sign language interpreters can be used in such important situations as interviews, testing, orientation, and training. Telecommunication devices can be used to give the deaf employee access to the telephone. Co-workers of deaf persons can learn sign language or can communicate through written notes. Deaf workers should be given equal consideration in terms of training and promotion opportunities.

Testing Deaf Persons for Employment

Mary Anne Nester and Magda Colberg, U.S. Office of Personnel Management

There are extensive data from various testing programs which show the impact of prelingually deaf people's deficit in verbal language. Deaf people as a group tend to have lower-than-average performance on verbal tests and average or higher-than-average performance on non-verbal tests.

The two main areas of concern in the testing of deaf people are 1) accurate communication of test instructions and 2) testing of intended abilities when verbal tests are used. An individual testing situation is recommended so that test instructions can be communicated through gesture, demonstration, and written notes, if the examiner does not know sign language. The problem of verbal test content is a more serious problem, for which several solutions were tentatively offered in this presentation. The main goal should be to prevent the medium of testing (e.g., the verbal medium) from interfering with measurement of the job-related knowledge, skill, or ability. When the job-related ability is verbal ability, then the appropriateness of the test must be given careful consideration.

Review of Test Performance of Deaf Persons. First with regard to reading level, it is an often-heard generalization that the average reading level of deaf persons is at the fourth grade level. A nationwide study of the performance of deaf students on the Stanford Achievement Test (DiFrancesca, 1972) showed, indeed, that the average performance of 19-year-olds on the Paragraph Meaning subtest was at the fourth grade level. The average performance of the best 19-year-old students (those who were able to take the most advanced battery of the Stanford Achievement Test) was upper seventh grade level. It has been suggested that such standardized reading tests actually overestimate deaf persons' reading ability and that a deaf person's ability to use language is not comparable to that of a hearing person who attains the same reading score (Drury and Walter, 1979).

This poor verbal performance is the result of the deprivation of auditory language rather than of low intelligence. Nonverbal tests of intelligence typically show deaf persons scoring in the normal range, while verbal intelligence test generally show them to be at least one standard deviation below the mean (Myklebust, 1964). It is not surprising, then, to find that the Wechsler performance scales are the most widely used intelligence test for the deaf (Levine, 1974). It should be noted that there are many prelingually deaf people who are exceptional and who do develop good skills in verbal language.

A study performed at the U. S. Civil Service Commission (Stunkel, 1957) showed the impact of the verbal deficit on competitive testing programs. A sample of Gallaudet College juniors and seniors were given a college graduate level test battery and their results were compared to the performance of 200 nonhandicapped applicants. The deaf students performed significantly lower on the Vocabulary, Grammar, Reading Comprehension, and Arithmetic Reasoning subtests, i.e., on all the verbal subtests. The two groups were equal in score on Figure Analogies, but the deaf students were superior to the hearing applicants on the Letter Series subtest. Since the battery consisted of four verbal subtests and only two nonverbal subtests, the deaf students were unlikely to earn scores that would qualify them for competitive job placement.

There was another characteristic in which the deaf students' performance differed from the hearing applicants'. For the hearing applicants there were significant positive correlations between all pairs of subtests (15 correlations). For the deaf group there were only six significant intercorrelations, and the correlations between verbal and nonverbal subtests were near zero or negative. (The exception was a correlation of .54 between Arithmetic Reasoning and Letter Series for the deaf.) These data suggest that, while for the hearing some common factor influences performance on all subtests, for the deaf this is not true. For deaf persons as a group, performance tends to be blunted by the requirement to use verbal language as the medium for reasoning. This conclusion has formed the rationale for many of OPM's test modifications.

Data collected by the U.S. Department of Labor on the GATB show a similar pattern of performance for deaf persons. These data are reported in the 3rd section of the GATB Manual. The results are that the deaf subjects scored below the mean on the three GATB aptitudes which are measured by verbal tests--Aptitudes G (Intelligence), V (Verbal), and N (Numerical). They scored at or above the mean on the other aptitudes--Spatial, Form Perception, Clerical Perception, Motor Coordination, Finger Dexterity and Manual Dexterity--which are measured by nonverbal tests. The deaf subjects obtained consistently high mean scores on Aptitude P--Form Perception. Aptitude G was tested with a test battery referred to as "Nonreading G," which consists of Figure Series, Figure Classification, and Matrices, for 72 of the deaf subjects. The average Nonreading G score was approximately 15 points (3/4 of a standard deviation) higher than the average GATB G for these subjects.

Another set of data which show the difference between deaf people's performance on verbal and nonverbal tests is available for OPM's Professional and Administrative Career Examination (PACE), a test which was discontinued in September, 1981. A paper by Nester and Sapinkopf (1982) presents data on the test performance of deaf job applicants on versions of the test which were modified for the deaf. To characterize briefly the modifications, it could be said that the verbal content of the test was minimized as much as possible in the four sections of the test that were not intended to test verbal ability and that the difficulty levels of the items were kept at the same level as those on the regular test, based on data from hearing competitors. Deaf competitors scored below the hearing mean on the two subtests (Verbal Comprehension and judgment) which contained verbal question-types and above the hearing mean on the other three subtests.

Trends In Employment of Deaf White Collar Workers

Harry Zarin, Job Placement Counselor, Gallaudet College, Washington, D.C.

White collar employment is dependent to a great extent on advanced education. It is estimated that only 10-12% of deaf persons enter post secondary educational programs and that half of that group graduate from the programs. Therefore, a relatively small percentage of deaf adults will have the educational background usually required for white collar jobs.

It is only in the last decade that numerous post-secondary programs have become accessible to deaf persons. Until the late 1960's, Gallaudet College was the only fully organized program for deaf people. Now, however, deaf students have access to such institutional as the National Technical Institute for the Deaf (NTID), the California State University at Northridge (CSUN), and Delgado College, to name but a few.

In a recent large-scale survey of Gallaudet College alumni, 85% of those who responded reported that they were employed in professional, managerial, or technical jobs. The incomes of those who responded were comparable to the national average for college graduates. The majority were employed in jobs related to deafness, but other job areas included finance, real estate, the computer industry, manufacturing and transportation.

A recent survey in Changing Times showed that employers want to hire people with backgrounds in engineering, computer science, finance, accounting, and marketing. Only 13% of the employers surveyed were willing to interview students with a Liberal Arts background.

Future employment opportunities for deaf people will depend on a variety of things including the ability of deaf people and the educational programs they are involved in to keep up with the times. Deaf students must take advantage of educational opportunities, including part-time job internships and co-op experiences. They must also know their academic and communication skills and be able to communicate this information to employers.

PAPER SESSION

Behavioral Traits and Their Influence on Selection Procedures

Chair: Jeanne L. Kaida, Los Angeles Community College District

Influence of Work Groups on the Selection Process

James W. Fairfield-Sonn, University of Hartford

Everyone affected by a selection decision, whether to hire, to promote, to transfer, etc., expects to be satisfied with it. Dissatisfied employees tend to leave organizations and dissatisfied employers tend to replace those who cannot fulfill their needs.

A review of the literature on selection reveals most discussion of the process emphasize the need to find the best individual to fulfill an organization's needs. However, the matching process is often made more difficult by the fact work groups are also involved and their reaction to a new member may have a considerable influence on how satisfied the individual and the organization are with the process. This latter concern will be the focus of this paper, as it attempts to reveal the importance of considering another dimension in conceptualizing the selection process.

A work group should be considered in making a selection decision in the following situations. First, a work group might be performing a task that is not a primary function of the organization. Another situation occurs when an individual needs to be found to join a very cohesive group. A third situation of concern occurs when an individual will be placed in a work group which is hostile to the current and/or future goals of the organization.

The major work on organizational socialization has tended to focus on individual and/or organizational variables and glossed over the mediating influence of the work group in the process. Work on understanding group development has been done almost exclusively on intact groups and not on those undergoing membership changes or not focusing intensely on those changes.

This paper proposes a theoretical model for making predictions about group reactions to new members. A model consisting of two independent dimensions, called group receptivity and position placement, has been developed.

The basic assumption of the model is the entry of a new member into a group represents a significant intergroup as well as intragroup event. A group reaction to the new member will thus reflect both the group's receptivity, or openness, to a member of a particular group, as well as its reaction to having a particular position filled within the group.

The primary hypothesis of the model is the two dimensions, group receptivity and position placement, will combine in a pure case to create four general modes of group reaction to new members. These modes are:

Avoidance (closed group, peripheral position) - The group ignores the new member, little communication occurs between current group members and the new member, the group will continue its work without trying to include the new member.

Confrontation (closed group, central position) - The group attacks the new member, the new member's opinions are challenged, the new member is subject to insults and slander from current group members, and the anger of the group is displaced onto the new member.

Acceptance (open group, peripheral position) - The group is friendly towards the new member, group norms are freely shared with the new member, but the new member's opinions are not solicited nor is the new member encouraged to join in the group's decision making deliberations.

Nurturance (open group, central position) - The group warmly welcomes the new members, solicits the new member's opinions, excuses shortcomings in specific knowledge and skills, and group norms are freely shared with the individual.

Group Receptivity. Placing an individual in a group will create an intergroup event. Its character will be influenced by group boundaries, power differences, affective patterns, cognitive formations, including "distortions," and leadership behavior.

By examining each of these factors it should be possible to develop a prediction as to how closed or open a group will be to a new member. Boundary relations refers to the permeability of both physical and psychological group boundaries. Power differences refers to the type of resources which can be obtained and used by different groups. Affective patterns become influential to the extent positive or negative feelings are associated with each group. Cognitive formations including "distortions," will affect the group's perceptions of both objective and subjective experiences with the new member. The behavior of group leaders towards new members will reflect in revealing ways the boundary relations, power differences, affective patterns, and cognitive formations of the relations between the two groups. When all of these five factors are considered together it should then be possible to predict how closed or open a group will be to a new member.

The central notion of position placement is based on understanding dependency relationships. In this conceptualization, the most important bases of power are information and placement on decision making channels. Therefore, to evaluate a position's placement within a group it is necessary to assess the amount and kind of information routinely flowing through the position and the frequency and quality of decisions the incumbent would make compared to other group members.

When the two dimensions of the model are combined, they yield a matrix for predicting initial group reactions to new members. The four extreme modes of reaction would be avoidance, confrontation, acceptance, and nurturance.

Uses of the Model. Developing an idea of how a group is likely to react to a new member could be valuable information for the new member as he/she develops expectations about what kind of experience awaits them. The model could also be useful in monitoring the progress of affirmative action programs. Specifically, it could be used to detect patterns of selection decisions which are supportive of or in conflict with an organization's affirmative action goals. The model could be used as another performance appraisal measure. This could be achieved by tracking the history of an individual's relationships with a group over time.

Providing Feedback to Individuals: Why It Doesn't Always Change Behavior

Philip L. Quaglieri, Northern Illinois University

Research involving feedback has typically concentrated on behavioral consequences to it and paid little or no attention to the cognitive events preceding those behaviors. To understand the effects of feedback, we need information about the cognitive links (e.g., acceptance) preceding feedback's effects--the amount and nature of information that supports them, the evaluations that surround them, the level of commitment they imply. It is unlikely that any motivational or learning benefit would accrue from feedback *per se* unless the person first accepts the information. Therefore, it was hypothesized that feedback changes behavior to the extent that a person accepts it (i.e., perceives it to be useful in determining how well he/she is doing the job).

The research design has two parts. The first part corresponds to a traditional survey study and uses an information processing approach. Its purpose was to assess acceptance and the feedback dimensions that explain it. It is assumed that the way a person reacted to imagined job situations would reveal his/her idiosyncratic method for determining the usefulness of feedback (i.e., the acceptance model). To evaluate the effect of feedback acceptance upon behavior, an experiment was used. In this portion of the study, groups of participants were provided feedback tailored to their acceptance models and observations were made regarding their responses (change in learning) to it.

Survey. The results of the survey showed there were three groups of participants (with at least $N \geq 20$) who shared similar acceptance models. Group One is composed of participants who placed greatest emphasis on specificity and trustworthiness. Group Two used power and expertise to make their decisions. Group Three is composed of those who relied upon accuracy and attractiveness to make their decisions.

The Experiment. The task was a proofreading exercise with texts selected from a national news magazine. Participants were given 4 minutes to proofread a text. After reading each text, participants were given feedback about their performance. As implied by the research design, however, one-fourth of the participants who shared a common acceptance model were given feedback such that the significant dimensions of their acceptance model were highly present; another fourth were given the feedback dimensions in a moderately present form; a third group was given the dimensions in a low present form and a fourth group was assigned to control condition and received neutral feedback.

Chi-square analyses were used to examine the association between the feedback forms (high presence, moderate presence, low presence and control) and the direction of behavior following feedback intervention. It was expected that those receiving high presence feedback would be more likely to increase their scores following feedback than those receiving the other forms of feedback.

Results. The data indicate that the high presence conditions are more strongly associated with increases in proofreading accuracy than any other condition of the acceptance model. Further, increases in proofreading accuracy occur more frequently the closer a participant is to the high presence condition of his/her acceptance model.

Discussion. The results of this study support the hypothesis that feedback effects behavior to the extent that it is accepted. Those persons receiving feedback that were congruent with the high presence condition of their acceptance models performed better than those in other conditions of their acceptance models. Further, it was observed that in all three acceptance models, those receiving high and moderate presence feedback had higher proofreading scores over the experimental periods than those receiving low presence and control feedback.

The results provide considerable insight as to why feedback does not always elicit the expected or desired behavioral change. Simply, the feedback message was not acceptable (useful) to the recipient.

Leadership Effectiveness: A Management Perspective

Debbera A. Diehn, Troy State University

The current journalists in management science dwell upon the need for leaders; for the development of a coherent operational theory of administration. This study tested one of the contemporary contingency theories that has shown potential for predicting leadership effectiveness: the Situational Leadership Theory (SLT).

The major purpose of this research was to determine whether the perceptions by supervisors of leadership effectiveness for their subordinates co-varied with the congruence (or match) between the supervisors' leadership style and supervisors' task-relevant ability. The supervisors rated the subordinates on ability levels; subordinates rated the supervisors on leadership style and leadership effectiveness.

Data Source and Sampling Design. Data were collected by mail survey with 300 questionnaires being mailed to State Directors of Vocational Education and Head State Supervisors of the five traditional service areas (Agricultural Education, Business & Office Education, Home Economics Education, Industrial Arts Education, and Trade & Industrial Education). Two hundred forty-one (80.34 percent) surveys were returned and two hundred eighteen (74.00 percent) were usable surveys returned in time to be reported in this study. In addition to the initial mailing, two follow-up mailings were necessary to achieve this rate of return.

Findings and Conclusions. Three major research findings were evident regarding leadership effectiveness. The undergirding premise of the SLT is that leadership style results from the adaption of leadership style to subordinates' task-relevant ability. A comparison of the research results do not support this tenet of the theory. However, support is provided for the tenet that task-relevant ability had no prior or independent relationship to leadership effectiveness. Only qualified/partial support is provided to the theory proposition that leadership style must be matched with task-relevant ability to ensure optimum leadership effectiveness. Further, results of this research do not support the proposition of a significant interaction of leadership style and task-relevant ability.

Practical Implications. (1) There is no ONE best leadership style; all leadership styles were perceived as being effective depending upon the situation. (2) Generally, the leader who is "human" to subordinates, exhibits a high relationship leadership style, is perceived to be the most effective manager. People like to be treated like people!! (3) The leadership style of a manager is determined by the behaviors of the subordinates. If subordinates perform in mature, responsible and appropriate ways, the manager will be supportive or leave them alone. But if they do not produce and perform in responsible ways, subordinates know their boss will be "all over them." (4) Leadership style, when matched properly to the situation, is perceived to be the best indicator of leadership effectiveness. Those leaders who adapted their leadership style to the situation and to the ability levels of their subordinates were perceived to be the most effective.

PAPER SESSION

Assessment Centers: Research and Application

Chair: Lourdes A. Lacomba, South Carolina Interagency Merit System
Discussant: Janet L. McQuire, Arlington County, Virginia

The Assessment Center As Psychological Process: An Analysis and Recommendations

Mitchell S. Shack, Department of Management, Bowling Green State University and Peter Bycio, Department of Psychology, Bowling Green State University

In the application of the assessment center method, it is assumed that behavior crucial to job performance will be observable in the assessment center. It is also commonly believed that regardless of their background, all trained assessors will observe a broad spectrum of activity and employ the established dimensions to form a final judgment. Finally, it is assumed that the consensus procedure compensates for individual error and yields a more accurate assessment than any of the individual assessors would have made.

However, although a number of assessment centers have been defended successfully on the basis of content validity, very few studies have looked at the relationship between dimension ratings in the assessment center and dimension ratings on the job. Of direct interest to the present discussion are recent studies which have failed to demonstrate the discriminant validity of the assessment dimensions.

Studies by Sackett and Dreher (1982), Borman (1982), Turnage and Muchinsky (1982), as well as our own work, suggest that assessors may not be using the dimensions in the intended manner. Our work has involved data from three separate assessment programs which vary in number of exercises, and assessor background.

A confirmatory factor analysis was used to evaluate a multitrait-multimethod matrix of ratings from a one-day assessment center. Candidates were assessed on eight job-related dimensions such as "organizing," "decision making," and "flexibility." Confirmatory factor analysis algorithms permitted the determination of the proportions of "trait," "method," and "error" variance associated with each assessment center rating.

The analysis yielded three conclusions that are of direct relevance to this discussion:

(1) Certain types of job simulations yielded ratings that were of significantly higher quality than others. By a "higher quality" we mean that some of the ratings had higher proportions of valid trait variance and lower proportions of method and error variance relative to ratings from other exercises.

(2) Second, almost all of the ratings were most heavily influenced by method variance, i.e., evaluations of a given dimension changed substantially as a function of which job simulation was being used.

(3) There were also extremely high correlations among the ratings of the individual job dimensions. The implication is that assessors were not distinguishing between the behaviors that were relevant to the different assessment dimensions. We are particularly concerned that dimension ratings may be fed back to the candidates as part of a developmental program when these ratings may not be accurate reflections of the candidate's behavior. Evidence was presented which suggests that what is being assessed within many assessment centers may be a single overall quality. We would like to focus now on a few of the psychological processes that are likely to be involved in developing the final products of assessment. The products of the assessment center are nothing more than human judgments, based on information which has been processed and transformed by the human mind. The processes include the acquisition of information, its encoding or categorization and organization, as well as its storage, retrieval, integration and interpretation.

The capacity of the human information processing system is limited in its ability to integrate vast quantities of data simultaneously. Assessors must attend to and process some pieces of information and ignore others. Research extending into the area of social cognition indicates that, among other things, the characteristics of the stimulus person, the situational context, and characteristics of the perceiver can affect which information will receive attention and how it will be interpreted. Characteristics other than dimension relevant behavior may influence the assessment. If the same extraneous characteristics affected each of the assessors then a reliable and apparently valid assessment could result. The important point is that, although it is assumed that assessors are influenced by job related behaviors, it is now clear that as human information processors, they are prone to differentially attend, recall, or rate candidates who are salient for various reasons, including their race, gender, or attire.

When processing information about other people, individual expectations and preconceptions about the way the world works serve as selection criteria and guidelines, helping the perceiver to select information from the stimulus environment. These expectations and preconceptions help to organize the search for information as well as subsequent storage and recall. This suggests that the perceiver places novel discrete events into familiar categories. This kind of processing is often termed schematic processing. It is implicitly assumed that all of the various people who serve as assessors are able to learn to adopt the schematic structure which was developed for the assessment center through job analysis and "empirical verification." This is unlikely since assessors have had years to analyze, interpret, and organize their own experiences and typically only have a day or two to learn a different and artificial framework.

The impact of goals on the acquisition of information has implications for the assessment center method. Ideally, assessors should observe and note as large a sample of behavior as possible without being evaluative. Such an approach would minimize the observational biases while ensuring a large sampling of behavior. The observer's level of arousal may also have an impact on the accessibility of various knowledge structures and on the way in which information is processed. A wide variety of arousal states should narrow the focus of attention to the most salient cue.

Assessors often encounter an information overload situation which has the potential to be quite activating and stressful. Added to the extreme demands of observation and note taking, the assessors work in the presence of others, knowing that ultimately they will have to present their observations to this group of peers. The expectations of the presentation to come and the possible informal evaluation as well as the mere presence of other people may further evaluate the level of arousal of the individual assessor, thereby affecting his/her focus of attention and tendency to be influenced by personal knowledge structures.

In summary, the capacity of the individual's information processing system is limited and the perception of information is selective. Humans have developed heuristics to streamline the coding, organization and storage of information. The information overload characteristics of the assessment center can serve only to narrow the assessor's focus and cause her/him to employ potentially biasing heuristics to reduce the information to an amount that can be handled.

A component of successful assessment is appropriate rater selection. Persons appropriate for the assessor role can be identified partly through a carefully designed and carefully evaluated assessor training program.

Some aspects of the assessment task could be simplified. Check lists of the more common behaviors which candidates normally exhibit or forget to perform during an exercise could be used. The accuracy of observation would increase, and the task would be less exhausting. We are also investigating the efficacy of having one person observe and record behavior, and another (perhaps an experienced, succesful job incumbent) evaluate it.

A final modification involves replacing the face-to-face group concensus meeting, with a procedure in which written summaries of information are passed among the assessors for their comment and evaluation. This would eliminate some of the negative aspects of group process such as the biasing influence of dominant individuals in the face-to-face concensus process.

An Examination of Internal Assessment Center Processes for Compliance with the Uniform Guidelines

Craig J. Russell, University of Pittsburg

The purpose of this paper is to take a step in modeling the cognitive processes assessors use when they make subjective judgments (i.e., ratings) about the assessee. Since one of the testing procedures used in assessment centers is the clinical judgment of assessors, it would seem reasonable to evaluate that judgment for bias in the same way a "test" is. In this case, the "test" score from an assessment center is the overall assessment rating (OAR). Consequently, two ways to examine an assessment center for bias is to 1., examine the regressions of some job performance criterion onto the OAR for different subgroups or 2., examine the regression of the OAR onto various dimensional ratings made from the exercises and simulations.

Method. Data were obtained on 2191 assessees evaluated in a one-day assessment center for an entry-level managerial position. Seventy-one assessors evaluated these assessees between December, 1979 and March, 1981. Assessors were second level supervisors who had gone through a three-week training program in the observation and rating of behaviors. The assessors worked in teams of four.

The Center. Each assessee was observed in four situational exercises. Sixteen dimensions of behavior were evaluated on a 1 to 5 point scale by each assessor. Assessors were trained to view the dimensional ratings as forming four categories (Personal Qualities, Interpersonal Skills, Problem Solving Skills, and Communication Skills) in arriving at the OAR.

Analyses. Three sets of analyses were performed to 1) determine how well the OAR can be predicted by the eighteen dimensions, 2) confirm the a priori four factor groupings of dimensions (i.e., Personal Qualities, Interpersonal Skills, Problem Solving Skills, and Communication Skills), and 3) examine how the factor groupings of dimensions, (both a priori and empirically derived) are related to the OAR.

Results. The highest simple correlation with the OAR is Interpreting Information ($r=.69$). All of the correlations are significantly different from zero. In the multiple regression study eighty-one percent of the variance explained on the OAR by all eighteen dimensions are explained by two dimensions, ninety-two percent is explained by five dimensions. Visual inspection of the factor loadings from the common factor analyses forced onto four factors failed to confirm the a priori four categories of Personal Qualities, Interpersonal Skills, Problem Solving Skills, and Communication Skills.

Studies to examine the moderating effect of sex of assessee indicate that males tend to receive higher overall assessment ratings then females.

Predictive Power. Over 80% of the variance in the OAR is accounted for by the consensus ratings on the eighteen dimensions. One explanation would seem to be the simple fact that the assessors were trained to arrive at the OAR through the evaluation of the dimensional ratings. Over 92% of the variance accounted for by all eighteen dimensions is accounted for by only five dimensions. These five dimensions happen to use only the In-Basket, Case Proposal, and Group Meeting assessment exercises as primary sources of information. A factor structures analyses were reported with interpretations of the factor loadings.

Conclusion. There is one major implication for the actual conduct of assessment centers in the future that comes out of this study. The way to ensure that the information gathered in an assessment center is appropriately combined into an overall rating is to take the decision out of the hands of the assessors and, instead, use an arithmetic scoring rule.

There are at least two advantages in the substitution of a decision rule in place of assessors arriving at the OAR. First, human error in the integration of assessment center information to arrive at an OAR would all be eliminated. All that would remain would be the possibility of an arithmetic error on the part of the clerical worker who combined the ratings. This would eliminate one potential source of error for which the firm could be penalized under the law (e.g., sex, race, and age bias).

Second, it would decrease the cost of running the assessment center. The less costly clerical worker's time would be substituted for that of the four assessors. This would amount to approximately fifteen to twenty minutes per assessee in the current assessment center.

Consultant-Agency Cooperation in Conducting Research On A Promotional Assessment Center for Police Lieutenant

Dennis A. Joiner, Dennis A. Joiner and Associates, Sacramento, CA; and Phil Carlin, City of Tucson, Arizona

This paper presents a model for consultant-agency collaboration for professional research. It is based upon an assessment center examination process utilized in the Tucson Police Department in Arizona.

One area of interest is the presenter's experience of candidate acceptance and reaction to a two phase candidate feedback procedure. The feedback process consisted of meeting with each candidate on a face-to-face interview by appointment basis followed by sending each candidate complete numerical and narrative documentation on their performance. The process allowed the authors to investigate participant reactions to the process and changes in their reactions through the use of questionnaires immediately after participation (before results were known) and again after the results were out and participant feedback had been provided.

Research was also conducted on the issue of integrated versus unintegrated data for determining final scores. This research was conducted by obtaining initial ratings on the skill factors or performance dimensions from each trained assessor prior to the integration session. The unintegrated score totals were then compared with the final integrated scores and overall consensus scores assigned for each candidate.

Development. Working directly from statistically derived job analysis results, supplemented by situational data and work samples, four exercises were developed which would simulate the most essential task areas in the classification.

For documentation purposes, matrixes were prepared which illustrated the relationship of (1) the essential tasks to the behavioral dimensions to be observed and assessed; (2) the essential tasks to the exercises to be utilized; and (3) the exercises to the behavioral dimensions to be observed.

The exercises developed included one leaderless group discussion, an informal oral presentation exercise, an in-basket exercise, and a written report exercise. Once the exercises had been developed, the consultant again visited Tucson. During this visit, all the developed exercise materials were discussed with the agency selection specialist and top Police Department management in relation to the supporting documentation from the job analysis.

Candidate orientation is a very important part of any assessment center examination process. The approach to candidate orientation was to send general information to candidates in written form. In addition to the written material, all candidates reported for a two-hour orientation session prior to their participation in the first examination test instrument or exercise.

Prior to the on-site assessor training, each assessor received a comprehensive package of pre-reading materials. Sending out the pre-reading material decreased the time necessary for on-site training to one full day.

During the process, each candidate was independently observed and evaluated by two different assessors in each exercise. The procedure for establishing candidate ranking was a combination of cumulative scores on each performance dimension with the addition of an overall evaluation score which the assessors established by consensus.

The total candidate group of 52 candidates required one full day of assessor training, four full days of assessment (13 candidates per day) and two days of post-assessment evaluation of candidate performance by the assessors to develop the final ranked list.

Results. The consensus of the assessor team was that the top 25 of the 52 participating candidates on the score ordered list had demonstrated sufficient skills to be considered job ready and placed on the eligible list. No protests or appeals were voiced by the candidates regarding the content, methods or procedures used in the examination process.

Candidate Feedback. Methods for providing performance feedback typically range from simply providing the one score which reflects the participant's rank on the eligible list to oral or written summaries of the participant's overall effectiveness by exercise or performance dimensions.

The approach for candidate feedback consisted of the agency selection specialist, who was thoroughly involved in the process, providing face-to-face feedback in the form of oral summaries of the narrative comments supplemented by specific quotations from the actual rating forms. Of the 52 candidates, 48 voluntarily participated in these 30-minute feedback sessions. Providing feedback face-to-face directly from the rating documents seemed much more effective than providing wirtten summaries of the information. No adverse reactions of any kind were perceived by the agency selection specialist. To the contrary, the participants seemed to appreciate the openness and honesty of the feedback procedures.

Research Conducted. During the job analysis phase of the project, the authors decided to conduct research on two aspects of the process: 1) participant reactions before and after performance feedback, and 2) an investigation of integrated versus unintegrated or clinical versus statistical combination of scores. To test the reaction of participants to the assessment process, each candidate was asked to complete a questionnaire at the conclusion of their participation in the exercises. To test the hypothesis that the responses would be slightly less positive after the results of the examination were out and feedback provided, a model for matching forms for comparison was developed. While the individual ratings assigned by the uninformed candidates averaged in the high positive range (good to very good or effective to very effective), the ratings assigned after the results were out average in the moderately positive range (good to fair or effective to neither effective or ineffective).

Weeks after the assessors had met over a two-day period and discussed and finalized all scores for all dimensions and assigned overall consensus scores for all candidates, the initial ratings were compiled without the overall consensus scores. Three different correlations were then computed. These correlations indicate that there might have been a difference in the rank order list if the overall consensus score had been the only score used for ranking candidates.

A comparison of the pre-integration dimension scores and the post-integration scores (including the overall score) produced a correlation coefficient of .983771. The rank order lists produced by these data are also identical suggesting that at least in this case, the integration session made no difference in the final results of the examination.

Combining Multiple-Class and Class-Specific Assessment Center Exercises For Time and Cost Savings

Jurutha D. Brown, City of Los Angeles

Since approximately 1977, an increasing number of management level examinations in the City of Los Angeles have included an In-Basket exercise, referred to in examination announcements as a Management Exercise, as a portion of the overall competitive examination. The use of a full assessment center as the competitive examination for managerial or executive level classifications in City service is not possible because of various financial and practical considerations. The Personnel Department believes that the use of management exercises or portions of an assessment center as part of the civil service examination is the best available method for evaluating managerial skills, traits and abilities.

For four high level, engineering oriented, managerial classes (City Engineer, Chief Deputy City Engineer, Deputy City Engineer and Assistant Director Bureau of Sanitation) the Personnel Department staff recommended the use of a single (the same) management exercise, designed and administered as a qualifying and weighted portion of all four examinations. Staff's recommendation was that the common management exercise be used as one weighted portion of separate civil service examinations in conjunction with other tests or exercises that would be specifically developed for each classification to measure critical knowledges, skills, and abilities not addressed by the management exercise. Staff believes that one or more professionally developed management exercises specifically designed to evaluate a group of dimensions can very appropriately be used for more than one class if the same group of dimensions is critical to all of the classes for which the exercise is to be used.

Other bases of staff recommendations were related to the practical and financial factors which must be considered when several related management exercises are to be conducted over a relatively short period of time involving the four positions. The practical question involves the advantages which may accrue to a person who may be applying for more than one of the four positions and be benefitted by the required feed-back on the first examination (taken for the first position applied for). Since the amount of the effect of these external factors cannot be specifically determined and factored out of the score, staff recommended that all potential candidates for a group of related examinations be administered the management exercise at once, prior to the announcement of any individual examinations. Then, when individual examinations were announced, the management exercise scores of any combination of candidates would have been achieved at the same time, under the same conditions.

The average cost to the City for the development and scoring of a single In-Basket exercise by a consultant has been approximately $5000. This does not include the cost for the City staff time involved in coordinating and administering the In-Basket process and developing and administering the remaining parts of the examination. If the cost of job analysis, In-Basket development and scoring for a multiple class exercise were as high as twice the cost for a single exercise, the use of the exercise for four classes would result in a savings to the City of at least $10,000 or 50%. In an era of declining budgetary resources, such a savings lends further support to the use of a multiple-choice exercise.

Engineering Executive Management Exercise - Multiple Class Assessment. A job analysis was conducted on each of the four classes involved in the recommendation, and an In-Basket exercise was developed. The job analysis phase of the project included a review of written information from Personnel Department files, analysis of questionnaire and interview data, and work samples from incumbents and supervisors, ratings and rankings of tasks and knowledges. A 31 item In-Basket which simulated important tasks common to all four Engineering Executive classifications was developed to measure written communication, ability to analyze and solve problems, planning and organization skills and management control.

Class-Specific Assessment. In addition to dimensions for measurement in the management exercise, the job analysis identified other knowledges, skills, abilities and traits critical to successful performance in each of the four subject classifications. The previous Civil Service Commission action established a weight of 35% for the Engineering Executive Management Exercise. Based on the job analysis and the examination history for each of the four classes, the remaining portions of the examinations were developed and announced. Certain dimensions were evaluated in more than one exercise in each examination. Such multiple considerations of important job elements, while not as rigorous as a full assessment center, is more desirable than the single consideration and evaluation provided by the traditional 100% oral examination.

Results. A total of 23 candidates participated in the Engineering Executive Management Exercise in October 1982. Of these candidates, 17 subsequently filed applications for and competed in one or more of the four civil service examinations conducted between December 1982 and May 1983. Although the concept of a common exercise was new, and therefore suspect to candidates and managers at the City of Los Angeles, the process was extremely well-received by both groups.

Conclusions. The Engineering Executive Management Exercise project was initiated as a compromise between separate class-specific assessment centers, which were desirable but not feasible, and class-specific oral examinations, which were feasible but not desirable. The City of Los Angeles' experience with combining multiple-class exercises with class-specific exercises has resulted in economies in the following areas: (1) financial resources devoted to development and scoring of exercises, (2) staff time devoted to coordinating and administering exercises, (3) candidate time devoted to participating in exercises.

SYMPOSIUM

Development of a Valid Selection Procedure for Nineteen Professional Classes in a State Merit System

Chair: William Rowe, Louisiana State Department of Civil Service
Presenters: Herbert W. Campbell, Richard E. O'Grady, and Richard H. McKillip, Psychological Services, Inc., Washington, D.C.

Project Background

William Rowe, Louisiana State Department of Civil Service

In 1980, the Governor of Louisiana retained the services of a labor law firm to respond to any inquiries into hiring practices in the state. One of the recommendations of the law firm was that the state devote more resources to test development and validation.

We obtained funding for four additional positions and for consulting services. Our first thoughts were to reduce our backlog and complete as many projects as money could buy. We even went so far as to write a Request for Proposal (RFP) with this in mind. We felt that in three or four years with the increased staff and the projects completed by a consultant we would have some validity evidence for all of our tests which had moderate testing and hiring activity. Prior to issuing our first RFP, we began to have second thoughts:

A. We had developed a number of tests for our entry professional classes that we were hoping to defend on a content validity basis. The tests were generally a combination of reading and data analysis developed from work samples.

B. It was our feeling that although these tests were more defensible than our previous tests, they were resulting in less effective selection. We had replaced a number of difficult abilities tests with some substantially less difficult abilities tests.

C. An outside party reviewed our proposal, lost sleep over it, and predicted a low bid of about $10 million by an irresponsible consultant.

A member of our staff suggested that, rather than simply complete more projects, we use this opportunity to complete a project that would make a significant contribution to selection in Louisiana. It was and still is our belief that the entry professional classes offered the most potential. In 1981 - 1982, we issued approximately 17,000 grade notices for these classes covering 2,500 positions and almost 1,000 appointments.

We then wrote a new RFP for the implementation of a test battery for the selection of entry professionals. We assigned three professionals and a clerk to work on the project, using additional staff as needed. The deliverables included:

1. Job Analysis Report
2. Survey of Options Report
3. Validation Report
4. Testing Manual

The biggest surprise to us was simply managing a project of this scope. For a number of the classes, we were attempting to use 100% of the incumbents in each of a number of phases. Coordinating the administration of hundreds of sessions to fill out questionnaires, pre-test, test, administer criterion measures, etc., in dozens of locations across the state was a nightmare. For the grading of the work samples, we needed high level supervisors for as long as a week. With all of the logistics problems, the encouraging parts were the tremendous cooperation and support from state agencies and employees and the willingness of our staff to make whatever effort was required.

Job Analysis

Herbert W. Campbell, Psychological Services, Inc.

As the first step in this project, we conducted a comprehensive job analysis of the targeted jobs. The object of the job analysis was to provide information for the following purposes:

1. to identify important job activities and critical or important work behaviors

2. to identify competencies or knowledge, skills, and abilities required to perform these critical or important work behaviors

3. to cluster jobs into homogeneous groups for purposes of test validation

4. to assess the content of existing tests for appropriateness

5. to serve as a basis for criterion development

The job analysis encompassed all positions within the 19 job classes to which employees may be promoted without taking a promotional examination. The titles and job codes of 46 professional position classifications were included.

The job analysis consisted of the following steps:

1. Generation of preliminary activity and competency lists

2. Review meetings with subject matter experts

3. Development and administration of the job analysis questionnaire

4. Analysis of the activity and competency ratings

5. Cluster analysis of position classifications

Generation of Preliminary Activity and Competency Lists. The job analysis procedure began with a review of existing class specifications, job analysis reports, and other relevant information. Task statements for the 19 professional job classes were compared, and a list of 38 activities was developed that encompassed similar tasks from several jobs within a single statement. Twenty-three of the activities were linked to two or more job classes and the remaining 15 activities each to one job class. All but one of these unique activities were linked to the Computer Programmer/Analyst job. Activity statements were edited as needed.

On the basis of the job class task lists and a review of relevant literature sources, a list of 16 competencies considered to be necessary for acceptable performance of the job activities was composed.

Review Meetings with Subject Matter Experts (SME's). Meetings were held with incumbents from the 19 job classes included in the study. SME's reviewed the preliminary lists of activities and competencies and suggested needed changes, additions, or deletions. SME's linked the task statements for their jobs to the general activities and determined which of the 16 competencies were necessary to perform the activities they had indicated were part of their jobs.

A representative sample of employees in 45 of the targeted jobs was selected to participate in these meetings. Human Service Worker III was later added to the list of jobs to be included in the study. The sample was stratified on the basis of race and sex and an attempt was made to include incumbents from varied work settings and locations.

The list of identified activities was modified after the early meetings to incorporate the changes suggested by the SME's. This process of review and refinement by job experts resulted in the number of activities included in the final list being reduced to 29. The task-activity linkup ratings made by the SME's in the 19 job classes indicated that these 29 activities adequately covered the critical and important tasks comprising their jobs.

No changes in the list of competencies were required as a result of the subject matter expert review. The linkup ratings for the competencies and activities indicated that all of the competencies were considered by a majority of the SME's to be very important for performing at least one activity. Therefore, the competencies adequately covered the knowledge, skills, and abilities required to perform the important activities comprising the targeted jobs.

The Job Analysis Questionnaire. The activities and competencies were incorporated into a job analysis questionnaire that was administered to incumbents in the position classifications included in the study. The questionnaire consisted of three sections. Section I requested demographic information. Section II contained the activity statements. Respondents rated each activity (in relation to their own jobs) in terms of frequency of performance, relative amount of time spent on the activity, the importance of the activity, and the amount of training required for successful performance. Section III contained the 16 competency descriptions. Respondents rated each competency in terms of the degree to which it distinguishes between different levels of job performance, how much difference there is between new hires on it, its relative importance, and how soon each competency is required of new employees after entering the job.

Louisiana Civil Service personnel were trained in the administration of the job analysis questionnaire. This was done to minimize inappropriate responses or missing information. Most incumbents in the targeted jobs except Eligibility Worker I were included in the job analysis survey sample. Because of the large number of Eligibility Workers employed by the State, only those employees in this job class with odd-numbered social security numbers were surveyed. The job analysis questionnaire was administered from January through March, 1983. A total of 2,835 completed questionnaires was obtained.

Cluster Analysis. Mean activity importance ratings were first computed for each of the 45 position classifications resulting in the formulation of classification activity profiles. These classifications were then grouped into clusters by performing a hierarchical cluster analysis. It was determined that a four cluster solution was the optimal means for grouping the 45 position classifications. This decision was based upon a statistical criterion which indicates the number of clusters where the profile similarities of positions within the same clusters and the profile differences between positions in different job clusters will both be maximized. The number of position classifications in the four clusters ranged from two to seventeen.

The Search for an Appropriate Cognitive Abilities Test

Richard E. O'Grady, Psychological Services, Inc.

A major task of the Louisiana project was to identify a single cognitive abilities test that would be suitable for use across all nineteen job classes. PSI conducted a comprehensive survey to identify examining options and to evaluate their validity and adverse impact.

The survey was conducted in two stages. First, a questionnaire requesting information on selection procedures and validity evidence for the nineteen jobs was developed and sent to state, county, and city government agencies throughout the United States. Second, the International Personnel Management Association (IPMA), and major test publishers were contacted, and the relevant literature was examined to determine the availability of appropriate tests.

Questionnaire. 169 questionnaires requesting information on selection procedures and their related validity strategies were mailed. Seventy-three (43%) of the questionnaires were returned to Louisiana. Nine respondents indicated criterion-related validity for four of the nineteen occupations. The majority of the selection procedures are supported by content validity alone. Agencies indicating that they had conducted criterion related validity studies were contacted for information on the method and results of their studies.

The twelve agencies that indicated using ability tests for seven or more of the occupations were contacted to determine whether the occupations were being tested by a single battery of tests or by job specific examinations. Eleven of these agencies employ job specific examinations. The remaining government is currently using a single cognitive abilities test battery for eighteen of the nineteen occupations. The battery, however, is not currently supported by criterion-related validity. The test questions contained information peculiar to that State and so were not generalizable to Louisiana.

The frequency with which various types of selection procedures are employed by the seventy-three responding agencies was calculated. Training and experience evaluation was the most common procedure (81%), followed by job knowledge tests (67%), interview (63%), and multiple-choice ability tests (55%). Other types were rarely used. Hunter and Hunter (1982) studied the accumulated evidence for the validity of alternative selection procedures. They found that, for entry level jobs, the average validity estimate for training and experience ratings was .13; for interview, .14 and for ability tests, .53. Thus, apart from job knowledge tests, the use of which may be questionable at the entry level, the respondents to the questionnaire are using the least valid procedure most, and the most valid procedure least.

Review of Other Sources. Additional information about available aptitude tests was obtained from the IPMA test library, the U.S. Office of Personnel Management, Tests in Print (Buros, 1974), Vocational Tests and Reviews (Buros, 1975), major test publishers, and personal contacts. PSI examined the descriptions of the numerous aptitude tests identified through these sources to determine if they were targeted to the appropriate level, normed on appropriate populations, and if they contained appropriate question types.

For the most part, the IPMA and OPM tests are geared toward specific knowledge required by each occupation, and not designed to tap cognitive ability. Job knowledge tests were judged not to be appropriate for Louisiana's examining needs.

Hunter demonstrated that professional occupations are of a highly cognitive nature. In a study across 515 jobs, the cognitive factor of the aptitudes measured by the General Aptitude Test Battery (GATB) was found to have higher validity for professional occupations than other factors.

In identifying potential tests for the Louisiana battery, PSI focused on test item types that have the highest validity for predicting job performance in occupations requiring the most cognitive ability, that is, professional occupations. Several studies on test items in the Professional and Administrative Career Examination (PACE) found the following item types to have the highest validity: inference, tabular completion, reading comprehension, letter series, and quantitative reasoning. Using Tests in Print (Buros, 1974), PSI test files, and major test publisher's catalogues, PSI selected a number of cognitive ability tests that appeared to be most appropriate for testing professional occupations. Most of the tests were normed and validated on student populations and were inappropriate with regard to difficulty level. A number of other tests were designed and validated for blue-collar or for clerical occupations and were rejected for similar reasons. A few tests survived this screen, and additional information was obtained from publishers. In most cases the additional information did not support their use for Louisiana's purpose.

Of the commercially available tests, the Computer Programmer Aptitude Battery (CPAB), produced by Science Research Associates, appeared to have enough evidence to warrant its use for the Computer Programmer/Analyst positions in Louisiana. It has been validated through criterion-related studies for programmer occupations in a number of settings. The CPAB is tailored for the computer occupational area and would require additional sub-tests to make it fully applicable to all jobs in the Louisiana situation.

The General Aptitude Test Battery of the U.S. Employment Service has been thoroughly validated over a wide variety of occupations. The research of John Hunter has shown the test to be valid across the entire occupational spectrum. The GATB is used primarily by state employment service offices. Its use by Louisiana would have to be negotiated with the U.S. Employment Service. A common practice of the USES, when it does allow the GATB to be used by public employers, is to have the test administered by the state employment service offices.

Discussion. Of the tests identified in the survey of state and local governments and the review of other sources of test information, no single, fully appropriate battery could be found. The only state developed battery had several weaknesses that precluded recommending its use.

The literature demonstrates that certain question types tend to be more useful than others for employee selection. In federal jobs equivalent in level to the Louisiana jobs, tests of inference, tabular completion, reading comprehension, letter series, and quantitative reasoning have consistently high estimates of validity. A test battery consisting of these sub-tests would produce a higher composite validity coeffient. The Office of Personnel Management, however, does not release federal test materials for use in the states.

Test Development and Validation

Richard H. McKillip, Psychological Services, Inc.

At this point, there was a serious problem. No fully satisfactory, ready-made test could be found, but the need for an examining vehicle still existed for the Louisiana State Civil Service. PSI perceived that there would be some general usefulness for such an examining vehicle and proposed to develop a test appropriate for Louisiana's needs, with criterion-related validity on Louisiana's jobs.

The need was for a cognitive abilities test suitable for selection of personnel into highly cognitive, complex jobs. Previous research has suggested a number of appropriate item types for this kind of testing effort. We developed an examining plan calling for four question types: quantitative reasoning, tabular completion, inference, and reading comprehension. The examining plan called for writing 200 questions; that is, fifty in each of the four item types to be tried out.

The reading comprehension and quantitative reasoning questions are the usual kinds of reading and quantitative reasoning questions that appear in many cognitive tests. The tabular completion tests had some missing numbers in the table, which usually can be deduced by subtracting or adding from some of the other columns or rows. Inference is an interesting item type. In it, a set of premises are given and a conclusion presented. The test taker has to determine whether the conclusion necessarily follows from the premises; probably follows, but not necessarily; is indeterminable, that is cannot be determined from the premises; probably but not necessarily false; or necessarily false.

The test questions were tried out on groups of 152 to 174 people. An item analysis, and an item fairness analysis were conducted. The best mix of questions with high item-total correlations and appropriate difficulty levels were selected from among those questions that survived the fairness analysis.

For each of the subtests, a plot comparing the relative performance of minority group members and majority group members on each question was developed; that is, relative to their own group performance. The correlation of item difficulties between the minority and majority groups is very high, .90. However, there is one question that seems to deviate considerably from the regression line or major axis line. Question 15 turned out to be relatively much more difficult for blacks than it is for whites. Such questions are examined to determine reasons why the question is operating as it is. The item might be changed to eliminate the source of bias. Typically, though, there will be enough questions to adequately test the construct, and the deviant item is simply excluded. That is what happened in this case.

The final test consisted of 100 questions: 25 each in the areas of quantitative reasoning, tabular completion, inference, and reading comprehension.

One of the first considerations in identifying the occupations on which to validate the test was sample size. A sample size of 200 for adequate statistical power was needed. Another consideration in selecting occupations was the availability of research subjects. This effort required great interdepartmental cooperation. The amenability of the occupation to criterion development was also taken into consideration.

Three occupations were selected: Computer Programmer/Analyst, Eligibility Worker and Probation and Parole Specialist. Validation studies were conducted on these jobs.

Special ratings were developed for each of the three jobs, special work samples and work problems were devised with the aid of subject matter experts. For one of the occupations, eligibility worker, a job knowledge test was constructed and administered. An uncorrected coefficient of .44 was obtained for total entry-level professional test score against computer programmer/analyst composite criterion score. For eligibility workers, the correlation between the entry-level professional test and the special supervisor rating was .31; the correlation between the entry-level professional test and the job knowledge test was .60. These are uncorrected coefficients. Reliabilities have not yet been calculated on the supervisor rating or on the computer programmer work sample. However, the reliability of the eligibility worker job knowledge test was .69. Correcting the coefficient for unreliability in the criterion, the corrected correlation coefficient is .72, so the entry-level test is really predicting the job knowledge of eligibility workers. Much more work is needed in the validation studies, including a test fairness study, which will be done on eligibility workers.

The next step will be to link the validity results with the other occupations to which the test is to be applied. This will be done through the cluster analysis approach and through a further analysis of the similarity of work activities across jobs. Section 7B of the Uniform Guidelines on Employee Selection Procedures allows transportability across jobs when 1) there is a validity study done on a job which meets the Uniform Guidelines validation study requirements, 2) a test fairness study has been done, and 3) the job to which the validity is to be transported has "substantially the same major work behaviors." Activities of the jobs on which the studies were done will be compared to the activities on the jobs on which criterion-related studies were not done to see if substantially the same major work behaviors exist.

Finally, a utility analysis will be performed. This analysis will follow the method of Schmidt, Hunter, McKenzie, and Muldrow (1973). Utility of a selection procedure is measured by the change in the productivity that would occur from the use of the valid selection procedure. No utility calculations have been made as yet and will not be, until our validation work is done. However, certain estimates of utility have been made for the programmer/analyst occupation. Let us suppose that only 10 people were going to be selected off this examination. And let's suppose their average tenure was 10 years; that is, they would all stay on the job 10 years. The validity coefficient is .50. The standard deviation of productivity based on Schmidt, et al. procedure is $10,413. The selection ratio is estimated at .10; that is, the State will hire only 10% of the people who apply. These persons are going to be selected from the top down, and the cost of testing is $10.00 per applicant. When you throw all these numbers into the utility equation you end up with a figure of $9,127.40 per person per year. Now if you multiply that by the average tenure you get 91,274, and if you multiply that by the number you select, which is 10, you end up with a net utility of $912,740 just for selecting 10 programmer/analysts with a valid selection procedure. The resulting long range benefits of using valid selection procedures across a large number of jobs and a larger number of selectees are enormous.

SYMPOSIUM

Development of Job Related Valid Oral Board Examinations

Chair and Presenter: Clyde J. Lindley, Center for Psychological Service
Discussant: Charles B. Schultz, State of Washington

Development of Job Related Valid Oral Board Examinations

Clyde J. Lindley, Center for Psychological Service,
Washington, D.C.

Over the years, interviews (orals) have surpassed all other personnel selection-promotion procedures in the number of superlatives characterizing them.

They are the most ancient of selection and promotion techniques as we know them today.

They are the most possessed of face validity.

They are the most used.

They are the most widely applicable in terms of type of job - from laborer to vice president.

They are the most subject to bias.

They have been the least reliable.

They have been least often adequately validated.

They have been least challenged in court.

They are least likely to be discarded to the personnel junkheap.

Hopefully, this symposium will suggest some ways to preserve the good of these superlatives and discard the bad.

Oral Examinations must meet all the current "professional standards" for selection and promotion techniques. This means insuring job relatedness and establishment of validity. There also needs to be empirically determined evidences of reliability of ratings (both individually and in group combination). There may also be concern related to combining oral board ratings with other selection techniques and arriving at the optimal weighting of the components of the selection process most scientifically by a partial regression equation.

Let me comment upon some major things that the personnelist must consider in planning an oral board examination:

The place in the selection/promotion process. This implies a comparison with other techniques which can be considered; and if more than one is to be used, various things that have to do with the way the oral board examination is to function in the combination. If one of the selection techniques is a written test, it usually becomes the screening device, for various reasons - its objectivity, its time economy, perhaps its higher validity. The oral board examination is rarely the initial screening device among a number of selection/promotion techniques. There is little evidence to indicate that it should be. Theoretically and scientifically if several techniques are valid enough to be used, the rule should be for applicants, all take all.

Weighting of the oral board examination. Most of the considerations here assume that the selection data is in quantitative terms (in a continuous distribution of scores or ratings). A review of techniques was discussed.

Types of questions to be utilized in the oral board examination. It is most important that all questions be job related, and critical to high performance in the characteristic being evaluated. Oral board examinations can be planned to evaluate a limited number of characteristics or a "cover the waterfront" combination. Broad coverage oral board examinations are more likely to be appropriate for higher level administrative-supervisory jobs. Broad coverages, unless broken down into elements, in terms of questions or characteristics covered, are for the most part to be discouraged. What oral boards are planned to evaluate should be based directly upon job qualifications determined by an adequate job analysis. In the interests of greatest reliability in selection and promotion, interviews or oral board examinations should be used only for those qualifications which cannot be measured by more objective techniques as written tests and objective performance measures.

Structured interview or an unstructured one. Most often structure implies the use of a predetermined list of interview questions - the interviewer is bound to these instead of free to ask anything he/she pleases. Structured orals have the odds on their side for the things we look for indicating reliability and validity in employee decision techniques. Dangers of too highly structured exams were discussed.

One to one - interviewer - interviewee contacts or oral examining board of several members. Both have their places as defensible choices. Factors favoring several members include primarily: desirability of board input from a diversity of board members, lessened challenges of unfairness, and greater ease of involving a board in total participation in the oral board procedural development; their entering into a board training procedure which can vastly increase validity of the oral board examination and its acceptability up and down the line.

Should all information about the applicant/candidate be available to oral board examiners? There is no categorically defensible "Yes" or "No" answer. In most oral board examinations the answer is "No." Oral board examinations are for the evaluation of qualifying characteristics which should not be influenced by (one might say "contaminated by") knowledge of information which constitutes the basis of other independent evaluations - as ability shown in test scores, education and experience records, etc.

Steps in planning and conducting oral board examinations. Steps are presented as exemplary procedures from the presenter's experience.

1. Conduction of an adequate job performance evaluation, setting forth tasks and qualifications for performing the tasks.

2. Determination of qualifications that cannot be evaluated by more objective techniques.

3. Appointment (or other means of designating) the members of the oral examination board or boards that will conduct the examination.

4. Oral board acquisition of thorough knowledge of job for which oral board examinations are to be conducted.

5. Determination of oral board examination questions to be asked and response guides. Our oral boards for most examinations have included non-situational type questions, which encourage candidates to make varying degrees of expansive answers. It has been our feeling and experience that situational questions ("what would you do in such and such described job-related situation?") tend to produce ratings too dependent upon correctness of response or judgmental quality of response rather than manner of responding or exhibition of the personal characteristics of concern in an oral board evaluation. Our experience has indicated that the development of response guides with sample types of behaviorally anchored responses considered by the board as "outstanding," "adequate," and "below average," are helpful. We have been careful to emphasize that these are not to be used as knowledge checks, but only as aids in evaluating more personal characteristics of responding or communicating. The guides, in our use of them, have been developed by the oral board members.

6. Development of instructions for the oral board members and the candidate. These have been developed by the consultants with input from members of boards with which we have worked, and members of personnel departments. They give general instructions and information to facilitate smooth running of examinations and relaxed approach by candidates.

7. Training of board members and trial run of the developed oral board examination has been a very important step. This has been done in a role-playing procedure, the total procedure being tape-recorded and played back for discussion among board members, with participation of the consultants in initial working out of the process. Subsequent in-house carrying out of the procedure is the expected rule.

8. Actual carrying out of oral board examination with candidates. The presenters who follow will describe how they developed structured oral board examinations. The presentation from the Sheriff's Department reports procedures we developed for them.

Transfer Oral Board Procedures Used by the Marion County Sheriff's Department

Karen Hamilton, Deputy Sheriff, Indianapolis

The department had a morale problem about our transfer or lack of transfer procedures for deputies to be transferred from one division to another (i.e. road patrol). It was felt there was a need to establish a procedure which would be fair and impartial to all deputies who applied for transfer. With the assistance of the Center for Psychological Service, we established an oral examination board for transfer when a promotion is not involved.

First, we had to review the job requirements for the vacant position. From the job requirements we established the minimum qualifications a deputy must have before applying for transfer. The qualifications cannot require prior job experience. The vacancy is then posted throughout the department in memo form under the sheriff's or colonel's signature. This memo contains the following information: a brief job description, the minimum qualifications, the rank assigned to the position, the tentative board date and how to obtain their appointment times, the deadline for applying and where to turn in request, how long the eligibility list is good for and an offer to any deputy who has questions about the procedure to come to the personnel office prior to the actual interviews so we may answer any of their questions. The memo is posted for 10 to 12 days to ensure all who are interested have an opportunity to see the memo.

After the vacancy is posted, the board members are selected. We have a five-member board and the highest ranking officer is the chair of the board. We try to stay with officers from the division which has the vacancy and, whenever possible, try to have at least one black and one female officer on the board. Selection is done by the personnel section and the deputy chief of the division which has the vacancy and with the approval of the sheriff. Due to the number of oral boards and the extensive training each board member must have, we have established a list of trained board members.

Once the board has been selected, we meet with them and ask them to assist in writing questions and answer guidelines which will be used for the board. The questions cannot be situational or require any detailed prior job knowledge. Once the questions have been developed, the board members must be trained. An average training session lasts about 3 hours. During these training sessions the members read through the questions and choose those with which they are most comfortable to ask during the oral interview. Each board member is given the same set of written instructions. These instructions inform the board on a variety of things.

They are instructed to ask only those questions assigned to them and not participate in idle talk with the candidate. Each board member must rate the candidate after each response has been given and their ratings must be independent of the other board members. They are also informed that only the presentation made by the candidate may be considered when rating and that they must not let any prior personal feelings contaminate the rating. The board is instructed about the "halo effect" and are encouraged to rate high if answer is good or to rate low if answer is bad. Also included in the instructions is the introduction and preliminary warm-up questions asked by the chair to help relax the applicant. To help the board feel comfortable with the procedure, the members actually conduct a mock interview which is tape recorded and then played back to the board. During the play-back, any rating problems are pointed out.

After the training is completed, the actual interviews are conducted. Prior to interviewing, the candidate is asked to read a set of written instructions which explains all facets of this type of interview procedure. The interview itself takes about 15-20 minutes. The rating form used has a point span of 0 to 10. Each board member rates the candidate independently and immediately after the candidate's response to each question. After the interview, the five score sheets are picked up by a member of the personnel section. The score sheets are then added, totaled, and averaged to obtain one score for each candidate. Each candidate's score is placed in numerical order from highest to lowest. Our consultants then place the scores into three cacegories: best qualified, qualified and least qualified. The three groups are then alphabetized, the scores are removed and a memo listing all three groups are sent to the sheriff. The sheriff then selects an officer to be transferred from the best qualified group. Due to the number of vacancies which occur in some divisions we established an eligibility list for 120 days using those officers in the best qualified group.

We have been utilizing this procedure since January 1981 and have held 36 boards to date. This system has been very well accepted. To help the deputies understand the procedure, we have an observer from the personnel section sit in on all oral boards. After the interviews are over this observer answers any questions the deputies may have as to the procedure or performance. The administration likes this procedure because it offers them a lot of discretion as to who they can choose for transfer. They are not forced to select officers in numerical order.

Oral Interviews as a Rating Tool for Librarians

Randolph Hilton, Montgomery County, Maryland

The Montgomery County Government Department of Public Libraries employs a structured inteview for staff librarian promotional examinations because it has proved to be a valid, job-related, and economically feasible instrument which enjoys a high degree of applicant acceptance.

Five years ago, Montgomery County used structured interviews for almost all employment opportunities. Shortly thereafter, written examinations were administered. However, the response from applicants to the written tests were generally negative, because they felt that the job of librarian preponderantly involved oral communication with patrons.

Written examinations would certainly be more economical if promotional examinations were routinely given to large numbers of applicants. In Montgomery County, most libraries have different areas of specialization (e.g., the arts, business, science) and may serve patron groups with different needs. In addition, librarians in one job class may function in one of several capacities, such as children's librarian, reader's advisor, reference librarian, etc. There is, therefore, a need to tailor each examination to the sometimes unique responsibilities assigned to a position at a particular library. Promotional announcements usually elicit a response of approximately 6 - 12 applicants who can be interviewed and which is manageable over a two-day time span.

Once the Library Department sends a vacancy notification to the Personnel Office, a job analysis is scheduled. The job analysis, which is conducted by a personnelist, takes the form of an interview with the immediate supervisor of the vacancy and a designated representative from the Library Administrative Offices.

After the job analysis, the type of examination to be administered is generally defined by the personnelist and the library representative. Given the previously-stated criticality of oral communication skills to the job of librarian, the examination usually consists of an interview.

The position is announced in an "Employment Opportunities Bulletin." The announcement provides 1.) a description of the position, 2.) the minimum qualifications for the position, and 3.) the knowledges, skills, abilities, and personal traits to be measured by the examination.

Applicants are rated by a three-member Qualifications Appraisal Board which is comprised of two specialized subject matter experts in the areas of librarianship identified in the job analysis (e.g. reference work, reader's advisory, children's outreach) and a representative from the library administrative staff. Usually, one member of the Board is a minority.

The examination itself is developed by the Qualifications Appraisal Board after a review of the job analysis findings with the personnelist. To ensure content validity, direct linkage must be demonstrated between the knowledges, skills, abilities and personal traits identified in the job analysis and the elements of the examination to ensure content validity.

The examination construction process begins with a presentation, by the Qualifications Appraisal Board, of several job-specific problems which are examined for their relevance to the KSAP's identified in the job analysis. The problems may be used to create work simulations which are typical of those which would be encountered by a librarian working in the designated specialty area in question.

Examination components are weighted according to the results of the job analysis. Once the examination is developed, it is forwarded to the Personnel Office for an analysis of the relation of the examination components to the job analysis and a review of the examination for relative fairness.

Previous to the scheduled interviews, Qualifications Appraisal Board members are instructed to rate applicants independently. The Board also decides the type of response which is acceptable for each question in the interview. A comprehensive rating framework for each examination component is developed by Qualifications Appraisal Board members. The framework identifies key elements expected in applicant responses, and provides point values for each element. Although the rating guide cannot always be followed rigidly, it does serve as a means of maintaining consistency between raters. When appropriate, professional aids and book catalogues are made available to the applicant for a brief planning period immediately preceding the interview.

After the conclusion of the examination, ratings are forwarded to the Personnel Office where results are then analyzed and adjectival ratings assigned. The appointing authority, which in this case is the Library Director, is then free to select any candidate in the highest rating category on the eligible list, as long as affirmative action objectives are met.

SYMPOSIUM

Firefighter Selection: Non-Written Tests

Chair: Nancy E. Abrams, Consultant, Personnel Management and Measurement

Physical Performance Tests: Setting the Pass Point

Nancy E. Abrams, Personnel Management and Measurement

Ms. Abrams introduced the symposium by stating that standards are traditionally set in an arbitrary way by using judgment of test developers or users, normative 1/2 standard deviation above or below the mean, and review of many studies. The Uniform Guidelines states in Sec. 5H, "where cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force." Arbitrary pass points do not meet this standard. Physical performance tests usually have a pass point to ensure that candidates have physical capacity to do the job.

This study explores methods of setting standards on Physical Performance Tests which meet the Guidelines. One method is criterion-related and involves: test incumbents, job performance data, setting a pass point so that a maximum of poor performers fail and good performers pass. This method needs a sufficient number of job incumbents, a representative group of job incumbents, and a good measure of job performance.

Method 2 (Content) collects job analysis data which identifies required KSA's or tasks and gathers data on how incumbents perform these tasks or estimates of the minimum level needed on the job. This method leaves much judgment to the job analyst and test developer and requires SME's to articulate job standards.

Method 3 is a modification of criterion referenced methods used for written tests. Sueprvisors observe applicants taking tests and then are asked to predict job performance. A point is selected where a maximum number of applicants predicted to be poor performers fail and applicants expected to be good performers pass. This method was used in two instances: a BARS anchored with critical incidents of only physical aspects of job; incidents presented in random order to supervisors and instructed to select incident which would most likely describe expected job performance, and not given test scores - only observe. There was some rater resistance but they could perform task reliably except for applicants unlike job incumbents (females). Further research is being performed on this phenomenon.

Firefighter Physical Agility Examination City of Rochester - 1983

Emily Kochanowski, Rochester, N.Y. Civil Service

The physical agility examination consists of eight exercises which test for strength, endurance, and coordination. A minimum standard must be met for each of the eight exercises in order to pass. Additional (ranking) points will be given to candidates who surpass the minimum standards for the last six exercises.

A candidate who fails any one exercise will not be allowed to complete the remaining parts of the exam. Following are descriptions of the exercises and the standards for each. Minor adjustments may be made before the physical test date due to changes in equipment or administration.

Exercise 1 (Height/Reach). This is a simple exercise which will check a candidate's ability to reach and manipulate equipment on fire apparatus. Candidates will need a total reach of approximately 80-84 inches.

Exercise 2 (Ladder Climb). This exercise checks a candidate's freedom from fear of heights as well as his/her physical ability to climb a ladder.

Exercise 3 (Hose Coupling). This exercise tests for a candidate's eye-hand coordination and, to a minimal extent, his/her hand strength.

Exercise 4 (Weight Lift). This exercise tests a candidate's strength - particularly his/her arm strength.

Exercise 5 (Ladder Extension). This exercise tests a candidate's strength, particularly his/her upper body strength.

Exercise 6 (Tunnel Crawl). This exercise is designed to test for the ability to crawl while carrying or dragging a tool, spatial orientation, and the ability to crawl in confined, darkened spaces.

Exercise 7 (Roof/Balance). This exercise tests for a candidate's balance and his/her strength and balance while manipulating objects in awkward positions.

Exercise 8 (Hose/Dummy Drag). This exercise tests a candidate's strength in lifting, carrying and dragging objects and gives special emphasis to endurance.

Retests. When a candidate fails an exercise, he/she may request an immediate retest for that exercise in the presence of the test supervisor. If the candidate passes on this second attempt, however, he/she will receive a maximum stanine score of "1" for the exercise, regardless of speed or repetitions. In no case will a failing candidate be allowed a third retest. If a candidate encounters malfunctioning equipment which, in the judgment of the exam proctor or supervisor, adversely affects his/her performance in a given exercise, he/she may, at the exam supervisor's discretion, be allowed an immediate retest or a rescheduled test appointment based on the malfunction. No other retests will be allowed.

NOTE: The paper described the exercise completely and presented guides on passing/failing and/or seconds scores converted into stanines. Other data about the test included candidate performance and scores, and problems and adjustments.

Comments on Minneapolis Firefighter Exam

Lance W. Seberhagen, Seberhagen & Associates

Background. The development of a new entry-level Firefighter exam for the City of Minneapolis last year was made more interesting by the fact that 1.) the city had been to the Supreme Court and back regarding racial discrimination in its past Firefighter selection procedures in the case of Carter v. Gallagher, which began in 1970 and still has not been resolved today; 2.) the Firefighter exam was the second court-ordered exam in the case because the first exam, which was based largely on personality and interest measures, resulted in adverse impact and had some problems identifying qualified candidates; and 3.) all steps in the development and validation of the new Firefighter exam had to be reviewed and approved by a court-ordered Affirmative Action Review Committee and the court before the new exam could be used.

Project Objectives. The City Personnel Department decided in advance that the new Firefighter selection process would include a physical ability test of some kind, but were not sure what other selection procedures should be used. Therefore, the consultant's assignment was divided into two phases:

Phase I. Conduct a detailed job analysis of entry-level firefighter and recommend a design for the entire selection process;

Phase II. Develop the physical ability test and other selection procedures approved by the City.

Phase I: Job Analysis. A multi-method approach was used to collect the job analysis data by a literature review, field interviews, direct observations, questionnaire surveys, and management review.

The heart of the job analysis was a questionnaire survey of both firefighters and fire managers. Firefighters (N=199; 69% response) rated 1.) their relative frequency of performing 223 tasks; 2.) when they developed 110 KSA's (i.e., knowledges, skills, and abilities); and 3.) the estimated validity of 35 potential selection procedures for firefighter.

Fire Managers (N=18; 69% response) rated the same 223 tasks for importance, the same 110 KSA's for both importance and scoring (i.e., pass-fail or ranking), and the same 35 potential selection procedures for validity. The questionnaire data were analyzed by computer, using CODAP/80. After a management review, the final list of KSA's recommended for use in the selection process was reduced to 51.

The consultant team then matched the final set of KSA's with the list of 35 potential selection procedures. The job analysis report recommended up to 17 different components for the selection process in three different configurations. They City Personnel department, Fire Department, and Legal Department all reviewed these recommendations carefully and decided to limit the selection process essentially to affirmative recruitment, physical ability test (with ranking), reading comprehension (pass-fail) with remedial reading training offered to those who fail, existing medical exam, existing training assessment, and existing 6-month probation (City Charter requirement).

The City decided not to test for much in the way of mental abilities (e.g., arithmetic, reading, mechanical comprehension) and rely primarily upon physical ability (e.g., strength and agility) to screen its Firefighter candidates. The City felt that this approach would minimize adverse impact to resolve the court case, while still measuring enough KSA's to provide qualified Firefighters.

Physical Ability Test. The City also decided that the physical ability test should be capable of in-door administration and use a work sample approach based on content validity, rather than using more abstract exercises (e.g., bicycle ergometer) which would have to be justified through validity generalization. The project team with further input from the City developed the following exercises:

Ladder climb. Climb an 85-foot aerial ladder on ladder truck set up on main floor of Minneapolis Armory (pass-fail);
Fire hydrant assembly. Connect hoses (ranked on time);
Room search. Crawl blindfolded in a room for three minutes (pass-fail);
Pike pole. Put hook of pole through 10 holes in metal panel (ranked on time);
Dummy drag. Drag 120-pound dummy 50 feet (ranked on time);
Stair climb. Carry 65-pound hose bundle up 3 flights of stairs and then pull second hose bundle up to balcony of Armory (ranked on time);
Balance beam. Walk heel-to-toe down-and-back on a 12-foot beam while carrying an air tank (ranked on distance).

All subtests used for ranking are weighted equally, with raw scores converted to standard scores for combination into the final overall score. All exercises were pilot-tested. As a result of the pilot-test, the physical ability test was revised in final form and now is ready for use.

Discussion. The new Minneapolis Firefighter exam illustrates some of the many administrative, legal, and psychological factors which determine the final design of any selection process. In this case, the various tradeoffs are magnified due to the controversy surrounding the test. Test administration will be time-consuming and expensive because there will be very little pre-screening. Additionally, candidates who score highest on the physical ability test will be required only to pass a reading comprehension test before they may be certified for hiring by the Fire Department.

PAPER SESSION

The Effects of Recession, Labor Relations, and Reform of Public Agencies

Chair: Judy Young, Arkansas Merit System

Labor Relations, Collective Bargaining, and Performance Appraisal in the Federal Government Under the Civil Service Reform Act of 1978

Douglas M. McCabe, Georgetown University

Introduction. This paper analyzes the labor relations, collective bargaining, and human resources management issues due to the establishment of performance appraisal systems in federal government agencies and organizations having labor relations obligations.

From a labor relations standpoint and for agencies engaged in implementing local plans, the most significant development was the issuance of two decisions by the Federal Labor Relations Authority in which the Authority ruled that the substance of critical elements and standards is not negotiable but that much of the rest of the performance appraisal system is negotiable.

The field research consisted of 1.) examination of relevant documentation and other primary sources of information in the Labor Agreement Information Retrieval System (LAIRS) of the Office of Personnel Management; and 2.) intensive and extensive interviews with federal sector managers, labor representatives, mediators, and arbitrators intimately knowledgeable concerning the interrelationship between federal sector collective bargaining and performance appraisal.

The oral interview questions were designed to elicit both factual and attitudinal information about federal sector labor relations problems in relation to performance appraisal issues.

The Federal Sector Collective Bargaining Framework. A major problem is the very limited scope of bargainable subjects. In contrast with the relative simplicity of private sector negotiations, federal sector negotiators must feel their way almost blindly through a confusing (to both sides of the table) labyrinth of laws and bureaucratic rules and regulations. There is a need for both sides to give serious attention to the problems generated by the uniqueness of federal sector collective bargaining.

Labor Relations and Performance Appraisal: An Analysis. In March, 1979, the Office of Personnel Management issued an Agreement Content Evaluation which provided guidance on most aspects of negotiating performance appraisal systems.

Agreement Review. Several problems were encountered by the Office of Labor Management Relations in reviewing the agreements. First was the diversity of performance appraisal plans. Another problem concerned the length of performance appraisal clauses. None of the agreements attempted to make the substance of standards and elements grievable. Also, nearly all agreements called for a summary rating of the employee's performance. However, the procedures used for developing the summary ratings varies considerably from agency to agency. Nearly all agreements state that performance appraisal rating will be used as a basis for assigning training, awards, promotions, and within-grade increases. However, there is considerable variety in the protections afforded an employee whose performance is rated less than fully successful.

Conclusion. In conclusion, federal sector labor and management have been encountering problems in contract negotiations over the issue of performance appraisal. Some negotiations have taken an unusually long period of time. In any case, the final and key consideration in federal sector labor management relations is to distinguish between personal leadership and impersonal administration. Maximizing the former and minimizing the latter reduce the need for employees to have recourse at the bargaining table.

Recessional Impact on Local Government Human Resource Management Systems: A Survey and Predictive Model

Kevin G. Love, Central Michigan University

Local governments are service oriented employee intensive organizations. The efficiency of local government HRM systems will provide the key in determining how well local governments can adapt to changing recessionary conditions.

Defining Local Government Environmental Factors. The overall environment within which the local government must operate its personnel functions is composed of three variables: population of the city/village which must be served, number of local government employees, and number of recognized bargaining units. Since only one of these variables directly dictates personnel system composition, it is proposed that the more bargaining units active within a local government, the more rigid the HRM systems will be within that organization. This rigidity should result in a greater inability to cope with recessionary conditions through changing personnel systems.

An Investigation of Recessional Impact. In an investigation of the impact of economic recessionary conditions upon local governments, the State of Michigan represents an extreme case. Due to the economic devastation of the automobile industry, the State of Michigan has been more severely affected by recessionary conditions than other states (Kozlowski, 1981). Economic conditions within Michigan (as indexed by unemployment, decreasing state revenues, etc.) are generally regarded as the worst within the fifty states (Verway, 1980). Therefore, the State of Michigan represents an appropriate target for an investigation of the impact of the recession upon local government HRM practices.

Survey Methodology. Each city and village government within the State of Michigan was asked to participate in the survey, a total of 532 local governments. Within each city/village the person in charge of HRM systems was asked to complete the survey questionnaire. Of the initial sample, 250 survey instruments were returned, a return rate of 47 percent.

The average number of employees working within the local governments was 199.6 with these employees belonging to an average of 1.6 recognized bargaining units. Of the 238 usable survey instruments, 155 were returned from a city/village with under 5000 population (65.4%); 34 from a jurisdiction serving 5,000 to 10,000 citizens (14.3%), 26 from localities with 10,001 to 25,000 persons (11.0%); 11 respondents from cities/villages with between 25,001 to 50,000 population (4.6%); 7 indicated populations ranging between 50,001 and 100,000 (3.0%); of indicated population between 100,001 to 600,000 (1.3%); and with a population of 600,001 or greater, one city/village was represented (.4%).

Survey Instrument. In each city/village in Michigan the person in charge of HRM was asked to complete a questionnaire. The survey questionnaire initially requested the following environment variables: number of employees and degree of unionization. The population of each city/village was indicated on the survey instrument after it was returned using 1980 census data.

The questionnaire requested ratings which indicated the impact of the economic recession on 25 human resource management areas. These areas were: communicating organizational goals; worker interest in organizational goals; worker acceptance of assigned task; delegation of authority; classification; wages and salaries; identification of future labor pool; projecting changes in work; recruiting; selection; placement; training; career counseling; organizational change; job satisfaction; supervision; use of authority; motivation of workers; worker perception of job and organization; performance level; analyzing worker skills/abilities; transfers; demotions; promotions; turnover.

Results. The areas of greatest recessional impact were: providing appropriate wages or salaries; identification of future workers and their availability for employment; projecting changes in what workers will do in the future; selection of new workers; providing orientation training and required skill training to workers; ability to change things within the organization with little disruption and maximum communication to workers; the job satisfaction of workers; and worker perceptions of the organizational environment.

Discussion. The impact of the economic recession on HRM systems within cities/villages throughout Michigan is not severe. As reported through the open-ended comments, many respondents indicated that the recession has had a positive effect in many HRM areas. For example, employees are more motivated to perform and less troublesome due to a fear of losing their job. This type of comment reflects a "management by intimidation" approach.

Index of susceptability. Based on the study results it is possible to create an "index of susceptability" for use in determining probable impact of the recession on various HRM systems. Using the degree to which each environmental factor predicts the recessionary impact (i.e., Beta Weights), multiplying each factor by the respective numerical value (weight), summing across factors, an index is created. An index of this kind provides a quantifiable and comparable measure indicating probable impact of the recession on specific HRM areas.

Recommendations. Although the reported impact of the recession on HRM areas is not severe, in the near future the situation may drastically change. With the advent of proposed "federalism." legislation responsibility for many state and federal programs will fall to the local government. This will place these organizations in the position of providing increased services to the public under decreasing budgets and staff.

Thus, it is important that actions be taken as soon as possible to develop a coping strategy to handle economic recessionary conditions. HRM systems should be reviewed in terms of their susceptability to negative impact by recessional conditions. Unions need to be aware of the negative impact of constraining HRM system flexibility. And finally, city/village management should not rely on the threat of layoffs as a motivational tool.

SYMPOSIUM

Designing and Implementing Personnel Functions in the 80's: The Critical Linkage of Personnel and Organizational Development Technologies

Chair: J. Kevin Ford, Michigan State University
Discussant: Ronald A. Ash, University of Kansas

The Critical Linkage of Personnel and Organizational Development Technologies

Sandra J. Miller, University of Tennessee

Despite well-developed technologies researchers, and practioners alike have suffered from failure of their programs to have an impact on organization effectiveness. This failure of personnel functions can often be attributed to an inability to attain employees' support in utilizing the technologies appropriately. Failure of personnel activities may be due to the restricted focus of the personnel manager's traditional perspective. We suggest changing the focus from an emphasis on the technique utilized to an emphasis on the intervention per se. The field of organization development (OD) concerns the social technology of interventions.

Initially there must exist a felt need for change by the employees. The second component of the OD perspective is the resistance to change. The third stage concerns utilizing OD techniques. The last component of the OD perspective is perhaps the most critical, user acceptance.

Resistance to Change. Six causes of resistance to change are: 1) the purpose of the change is not clear, 2) persons affected by the change are not involved in the planning, 3) fear of having to break or alter established social relationships, 4) communication regarding the change is poor, 5) a lack of respect or trust in the initiator of the change, and 6) a lack of top management support.

One of the underlying assumptions about change is that it occurs in phases. Kurt Lewin identified three phases: unfreezing, moving and refreezing. He proposed that without all the phases included in the planned intervention one could not trust the permanency of the change. A successful innovation involves unfreezing (if necessary) the present level, moving to the next level and refreezing group life on the new level.

Unfreezing corresponds to the development of a need for change. The moving phase would entail diagnosis of the problem, examination of alternative routes and goals, establishing the goals and intentions of actions and finally transforming intentions into actual change efforts. Refreezing involves the generalization and stabilization of the change.

Two Focuses of Organization Development. Two focuses in the field of OD which are intended to increase receptivity to change are focusing on the group as the basic unit of change rather than the individual, and the use of participation to facilitate acceptance of innovations in organizations.

Two of the most important advantages of participation are that it gives employees greater knowledge of what they are to do - they have a better understanding of the task, and it provides for greater user acceptance and commitment to the change.

Four OD Techniques. Two techniques that involve the use of groups are survey feedback and team training. The other two OD techniques are the nominal group technique and the delphi technique of decision making. The nominal group technique is basically a four step process: 1) silent judgments are made by individuals in the presence of the group, 2) in a round robin fashion, judgments are presented and listed on a flip chart without group discussion, 3) group discussion of each judgment for clarification and evaluation, and 4) individual reconsideration of judgments and mathematical combination to form a group response. The delphi technique is somewhat similiar except that the individuals are not together as a group.

Project Unity: Using OD Techniques for Implementation, Development and Capturing Employees Perceptions of a Performance Appraisal System

Raymond A. Noe, Michigan State University

A performance appraisal system which is developed with the direct involvement and participation of employees will likely result in a minimal resistance to the system and increase chances of successful implementation. This paper describes the development and implementation of a performance appraisal system in a Sheriff's Department in the south central Michigan area.

The initial step of the project involved establishing whether or not the employees in the Sheriff's Department perceived a need for change or felt that problems currently existed which needed to be rectified. The Nominal Group Technique (NGT) was used to facilitate committee member participation and to reach consensus regarding the objectives of the project.

Based on the NGT (nominal group technique), the following two objectives with the highest priority were identified: (1) to increase the level of trust and communication between members of the department and (2) to improve the quality of performance throughout the department. The objectives of the project as determined by the steering committee and organization survey results suggested the need to develop and implement a personnel technology which would be conducive to improving job performance and result in increasing the communication between administrators, sergeants, and officers. Behavioral Observation Scales (BOS) would be the appraisal system which would best meet these objectives.

In order to gather the critical incidents needed to construct the BOS, small groups of employees were used to generate the incidents for each job. The use of the small groups for gathering the information necessary to construct the BOS presented the opportunity for the employees to clarify their own job expectations as well as to clarify the role of fellow employees in the Department.

The following two major benefits can result from the use of employee groups and OD techniques like NGT which increase employee involvement and participation.

1. Employees can clarify their own job expectation as well as gain an understanding of the work of fellow employees.

2. An increased probablility of acceptance and understanding of the appraisal system.

The Use of Survey Based Approaches to Examine and Reduce Trainee Resistance to a Police Weapons Training Program

J. Kevin Ford, Michigan State University

Training programs often fail to meet their stated objectives because various parameters are not addressed in the development and implementation of the program. One key parameter rarely addressed is the potential for resistance to the training program by incoming trainees and its impact on training success.

The focus of the presentation is on a skills oriented training program developed for a large midwestern city police department. A 20 hour training program was developed by the training staff to impart the skills necessary to become proficient in the use of the PR-24, an advanced police baton. The training program emphasized practiced instruction with the weapon.

OD specialists have warned that any major policy change, such as the one to standardize the intermediate weapons, will generate resistance from those affected by the change. The unique challenge posed by resistance calls for the utilization of techniques to reduce resistance by unfreezing and changing prior beliefs. This process of unfreezing and change should begin before the actual skills training program in order to increase the chances for maximal learning and performance. OD techniques which identify personal and normative beliefs are powerful tools for examining potential resistance to personnel technologies. Similarly, OD techniques are available which can aid the unfreezing and changing of current beliefs.

To identify the potential for resistance to the weapons training program, a survey was sent to the 167 (of the 207) recruits who had been given PR-24 training and were currently on patrol. Ninety-seven officers completed the survey which addressed the job relatedness of the training and the beliefs and attitudes of the officers towards the PR-24. The results of the survey indicated that the officers' attitude toward the weapon had become significantly less favorable and their confidence level had decreased since they had been trained. For the officers trained in the PR-24, the longer they had been on patrol, the less likely they were to carry the weapon. Regardless of patrol tenure, the trained officers universally perceived that the attitudes of their patrol sergeant and fellow patrol officers were highly unfavorable to the use of the weapon. The negative attitudes perceived by the trained officers from the survey strongly suggest the potential for resistance to the PR-24 training by the untrained patrol officers and sergeants.

As discussed in the OD literature, support from all levels of management is necessary for successful change efforts. The support of the patrol sergeants, as first line supervisors who closely interact with the officers, was seen as crucial to the success of the program. The action plan concentrated on changing the sergeants' beliefs and attitudes about the weapon. A survey feedback approach was chosen as the major technique for change as it has been found to be quite effective in unfreezing and changing existing beliefs. To examine the effectiveness of the survey feedback approach, a subset of 21 sergeants were combined as a group to complete the PR-24 training program. The training program was expanded from 20 - 32 hours to increase the amount of practiced instruction and to incorporate an attitude change component into the skills oriented training program.

A comparison of the pretraining and the posttraining survey demonstrated that the sergeants' attitude had become significantly more favorable towards the weapon. A plan of action which increased the involvement of the sergeants in the change process was developed and accepted by top management.

PAPER SESSION

New Developments in Selection for the Fire Service

Chair and Discussant: Richard H. McKillip Pyschological Services, Inc.

Validation of Fire Officer Promotion Procedures

Don Sommerfeld, Ph.D., Michigan Municipal League, Ann Arbor

The purpose of this project (1981-82) was to develop and validate selection and promotional procedures for fire service officers. Using an extensive job analysis the project produced data on the job requirements of all levels of fire service officers in jurisdictions throughout Michigan which became the basis for the new procedures.

A job analysis questionnaire was developed based largely on the National Professional Qualifications System produced by the Joint Council of National Fire Service Organizations. The questionnaires were used to collect data about three subjects: (1) the rank, title and organization of the fire officer responding; (2) tasks performed by the officer and (3) what knowledge, skills and abilities (KSAs) are required to perform the job tasks.

Over 300 fire service personnel, from 40 different Michigan fire departments, participated in the job analysis survey, evaluating the importance and time spent performing or supervising the tasks required and rating the amounts of the various KSAs required. After these data were collected, factor analysis was used to organize both tasks and KSAs in logical groups.

In a separate activity, a group of fire service professionals were asked to match each KSA with the tasks for which they were required.

As a result of these activities, 10 KSA factors were identified which contained only KSAs which are critical requirements of fire service officer positions and which are required to perform tasks which are critical parts of the job. These 10 factors provide the basis for the production of content-validated examinations and selection procedures. Exact contents of these examinations and procedures for any specific rank in any specific jurisdiction will vary according to the specific requirements of that job.

A final result was the development of 487 written exam questions, which passed a critical review by SME's and which covered 6 of the KSAs. Oral examination topic areas were designed primarily for four KSAs, but might also tap some KSAs for written examination questions.

Problems in Job-Related Measurement of Reading Ability

Sandra S. Payne, U.S. Office of Personnel Management

Reading comprehension is a requirement for satisfactory performance in many jobs, and consequently a test of reading ability is frequently used in job selection. To enhance the job-relatedness of the reading ability test, it seems logical to have it be as closely linked to actual job reading requirements as possible.

Over the past several years I have used and refined a set of procedures for measuring the reading level of the written materials used in the performance of a specific job and then developing a reading comprehension test that measures the same reading level. In several instances I have gone a step further, and have used actual job materials as the basis for the reading comprehension questions to be used on the test.

I reasoned that there could be no fairer or more job-related method of assessing reading ability for selection purposes. The link with the job would be so strong that the test could be supported on a content validity basis. More importantly, the test would be so clearly job-related that it would be rationally as well as psychometrically defensible. Such an obvious job linkage could conceivably affect the acceptance of the test as a fair and unbiased predictor of job success. These benefits would compensate for the fact that the procedures are somewhat time-consuming, and therefore costly, to follow. Unfortunately, in at least one project in which these procedures were used, the results were not as favorable as we expected.

I first applied these procedures when working with several other psychologists on a project to develop a new written selection examination for firefighters in the District of Columbia. Reading comprehension was one of six abilities measured in the examination.

After an extensive and thorough job analysis, the written materials used in job performance were identified by the DFCD. These materials were a combination of DFCD and national publications. Only materials actually necessary for entry-level job performance were named. A Flesch reading ease index was computed for each of the required written materials, and also for all of the materials combined. The format of the reading comprehension test item was a paragraph comprehension, multiple-choice format. For content, our choice was between using paragraphs taken from general subject-matter materials, or--with careful item writing--using paragraphs from actual job materials. We chose the latter. Each question could be answered by a careful reading of the paragraph. No special experience or training would be required or would help in choosing the correct answers.

We believed that the reading portion of the test, although measuring a general ability, was essentially content valid. This validity was confirmed by the results of the concurrent criterion-related validity study that was conducted before the test was used operationally. The correlation of the new written test with the job performance criteria was .61, and the correlation of the reading comprehension subtest with the criteria was .49--both correlations significant at the .01 level. In addition, comments received from the incumbent firefighters participating in the validity study were generally positive on all aspects of the test.

The test was finally given in late 1980 to over 900 applicants. The great majority of these applicants were black. Almost everyone who took the test passed. Based on the passing point statistics, there was no adverse impact from the test. However, a disproportionate number of whites appeared at the top of the register. Shortly after the first selections were made in 1981, a class action was filed, alleging discrimination in employment in the DCFD in violation of Title VII. The complaint alleged that an unvalidated test was used for hiring.

A review and analysis was made of the arguments before the hearing examiner. The plaintiffs contended that it was inappropriate for the reading comprehension test to have been related to the reading level of job materials which could possibly be written at a lower level. They also contended that the reading level of the materials could easily be rewritten, without affecting the content of the materials, and therefore the selection test did not need to measure reading ability at the more difficult level. Their arguments must have swayed the Hearing Examiner, because again, his findings of fact in the case concluded that the reading level of the selection test was unnecessarily high, and the use of a lower level would not have affected the validity of the test.

In looking back, in addition to our testimony that validity would likely have been affected if we did not test for the job as it existed, we could also have more thoroughly testified about the inherent difficulty of some subject matter. We also could have made more clear the fact that many of these materials are national in origin, and are the standard texts used by firefighters in fire departments across the country.

One measure of worth is, of course, validity--and we do have that--rather substantial validity at that. Our overall test had higher validity than the test which was previously in use (which also was highly valid, but was not job specific). We also found in our consideration of alternative selection devices that the new test was more valid than any viable alternatives. As a result of this hearing a closer look is taken at the process of developing the job-related reading test. I am still intuitively in favor of using actual job materials if possible.

F. E. A. T. Revisited and Restructured

Phil Carlin, City of Tucson

In 1979, under converging pressure from Women's Rights Organizations, Affirmative Action, Personnel, Fire Union Officials, and Fire Department Administrators, the First Encounters Acceptance Test (F.E.A.T.) was developed and administered to measure the physical abilities of firefighter applicants in Tucson, Arizona. Since that time, fifty-four departments from New York to San Francisco, have reviewed the F.E.A.T. concept. Briefly, F.E.A.T. stresses work task fidelity and requires candidates to complete tasks (events) in continuous sequence within a total established time.

F.E.A.T. requires an applicant to perform seven tasks related to firefighting work. All tasks must be completed in continuous sequence within a total established time. Before starting F.E.A.T. the applicant is fitted with a Scott Air Pak weighing thirty pounds. Helmets, gloves, and turn out coats are provided for applicant protection. Applicants receive training aided by a video-tape presentation prior to attempting F.E.A.T.

General Description of Test:

1. Hose Roll - Applicant rolls and unrolls 50 feet of 2-1/2 inch double jacket hose.

2. Hose Advance Event - Applicant alternately drags 100 feet of 2-1/2 inch double jacketed, uncharged, unrolled fire hose and connects/disconnects 2-1/2 inch smooth base fire nozzle.

3. Ladder Set-Up Event - A 14-foot aluminum wall ladder is removed from a fire truck, carried around the truck and returned to its original position.

4. Fire Extension Event - Moving through a two-foot by two-foot attic opening applicants must stoop and crawl along simulated rafters 16 inches apart.

5. Ventilation Event - While standing on a ground level pitched roof, applicant strikes target object 56 times using eight-pound sledge hammer to simulate axe. Target must be struck with forceful action similar to venting a roof.

6. High Rise Event - Applicant carries a 50 foot, 2-1/2 inch double-jacketed hose roll and a high rise kit from a fire truck to the training tower's fourth floor before returning down the stairs.

7. Citizen Assist Event - Applicant picks up a 100 pound dummy and darries it around training tower.

SYMPOSIUM

Computer Assisted Testing in Government Jobs

Chair: David C. Myers, Advanced Research Resources Organization
Presenters: Joanne Marshall-Mies and Mark Schemmer, Advanced Research Resources Organization

Discussant: Theodore S. Darany, San Bernardino County, California

Computer-Assisted Testing in Government and Industry

David C. Myers, Joanne Marshall-Mies, and Mark Schemmer

Many jobs in government and industry have become increasingly sophisticated and complex with technological advances increasing the load placed on the incumbent's perceptual-motor and cognitive functioning. The speed and accuracy with which information is perceived, encoded, stored, transferred, and compared; the speed with which memory is searched and accessed; and the speed with which decisions are made, are all critical to worker performance and productivity. There is a need to develop valid assessment instruments that focus on processes such as selective attention, time sharing, memory, spatial visualization, and comprehension as predictors of job success. Similarly, valid measures of perceptual-motor skills have long been considered important requirements in a wide variety of jobs. While tests of perceptual-motor abilities added significantly to the predictive power of basic written tests, they fell into disuse because of their low reliability and expense of the apparatus. The increasing cost of personnel training and employment as well as the development of reliable computerized testing stations has created a renewed interest in these potentially valuable personnel assessment methods. Furthermore, some research indicated that the differences between subgroups (e.g., minorities) may be less with performance-based tests than on traditional paper-and-pencil tests.

A major program of research at ARRO focuses on solving problems related to development and validation of computer-assisted tests which measure a broad range of cognitive and perceptual-motor abilities. These types of tests offer the opportunity to account for untapped variance in job performance and provide a cost-effective way to test large numbers of applicants. The ARRO staff has developed job analytic methods that determine the important skills and abilities required to perform jobs successfully and safely. In the symposium the testing apparatus and software which measure these important skills and abilities were presented. One recent effort involved a nation-wide study to validate computer-interactive tests for the electric power industry; and another study involved development of perceptual-motor tests for the military. These types of advances in the development of computer-based tests have implications for extending the technology to other jobs such as law enforcement, fire fighters, technicians, maintenance mechanics, air traffic controllers, equipment operators, computer operators, managers, and decision-makers.

PAPER SESSION

Biodata as a Public Sector Selection Strategy

Chair and Discussant: Brian O'Leary, U.S. Office of Personnel Management

Increasing the Odds for Producing Valid Biodata Instruments

Carol Bullock, U.S. Office of Personnel Management

This is a report on an interjurisdictional validation study of entry-level firefighter tests for three Maryland counties: Anne Arundel, Baltimore and Howard. A biographical inventory was chosen to measure the KSAO's developed through a job analysis. A review of the literature on weighted applications and biographical items and a review of numerous firefighter and other validity studies were completed. Success with this format in the clerical and corrections officer occupations led me to believe that it might work with firefighters, an occupation where no test except a job knowledge test had ever been very good at predicting job performance prior to 1978.

Items were selected and developed based on their relationships to the characteristics we were trying to measure. Every effort was made to ask job related questions. The items were both hard (factually verifiable) and soft (subjective and self-descriptive). All items were in a multiple choice format.

Validity Study Sampling. There was a need for at least 300 subjects, preferably more to empirically key a bio-inventory. Because 300 subjects are difficult to obtain in most occupations, if you are working with a single jurisdiction, multijurisdictional studies become an answer to the sample size problem as well as to the resource problem within the personnel organizations of cities, counties and states.

Again, because at least 300 subjects were needed to develop a biographical inventory and because of a pending law suit, my validity study had to be concurrent rather than predictive, that is, use incumbents rather than applicants.

Item Selection or Development. Since personality and attitude testing are such sensitive areas of employment testing, it behooves us to try to show that the biographical inventory is job related or valid from a content standpoint as well as from a criterion related standpoint. To avoid unreliability problems, attitudes, interests and personality characteristics should be clearly operationally defined when they are included in job analyses questionnaires. When actually writing or selecting items to measure each characteristic covered by the bio-inventory, you should obtain the assistance of subject matter experts.

A large number of items should be included in your initial bio-inventory to increase the chances that you will find some that differentiate incumbents and applicants. A variety of resources are available to assist you in item selection and development. The Catalog of Life History Items developed by Division 14 of APA in 1966 is a resource that may prove helpful. Finally, the items should be pre-tested with recently employed incumbents and/or applicants to evaluate their power to discriminate among persons in these groups.

Criterion Development. Since bio-inventories are keyed based on supervisory performance ratings or other criteria of job success, the job relatedness of the criterion becomes very important to the success of one's efforts to validate these instruments. I developed an 11 dimension supervisory performance appraisal instrument using previously scaled behavior examples that had been developed for a nationwide firefighter study.

It is important that the criterion instrument be as reliable as possible. Job relatedness and clarity probably insure reliability more than cleverly contrived efforts to avoid bias or halo. Supervisors cannot rate employees consistently or accurately if they do not understand the rating scales and how to use them properly.

Feasibility of a Content-Valid Biographical Questionnaire For the Selection of Municipal Police Officers

Walter G. Mann, U.S. Office of Personnel Management

Problem. There are many problems with biographical questionnaires. They have to be revalidated every three to five years. They also require a large number of employees, from 400 to 1000. Biographical questionnaires have traditionally been excluded from public sector examining. Some reasons for this are: lack of job relatedness, lack of face validity, invasion of privacy, and adverse impact.

Present Study. In spite of these problems, BQs deserve serious consideration because they are based on the sound premise that information about the past is the best predictor of the future. They have satisfactory validity; validity coefficients in other studies have averaged .37. In addition, a BQ, because it is an extension of the employment interview, is acceptable to management.

In the present study a content-valid BQ was developed and used to predict turnover. An initial step was to identify evidences in the applicant's background that indicate he/she has experienced situations similar to those in the job for which he/she is being examined. The rationale is that the more situations already experienced like those in the job applied for, the greater the motivation to stay on the new job; or with onerous aspects, the less motivation to leave the job.

If a BQ tapping past experiences is based on an analysis of the job and its requirements, it will be content valid and face valid, have *a priori* correct answers, and require but few employees to develop. Such an approach will also avoid problems of invasion of privacy, discrimination, and shotgun empiricism.

Method. In the job analysis phase of the larger study, a random sample of 121 police officers rated the importance of aspects in their present jobs. Eight items concerned verifiable information about the respondent's background. In addition, each item was based on a part of the police officer's job that was a potential source of trouble. Criterion measures were collected after 2 1/2 years on the job. Since the intent was to develop a measure of motivation to stay on the job, turnover was the key criterion.

Results and Discussion. The content-valid BQ correlated .33 with turnover ($p < .05$). The significant prediction of turnover is evidence of the BQ's predictive validity. The content-valid BQ predicted turnover over a 2 1/2 year period. It is possible therefore that content-valid BQs do not have to be revalidated as frequently as BQs developed using a criterion-related validation approach. Research is needed to study the deterioration in the predictive validity of content-valid BQs over time.

Limitations of the Present Study. Fakability was not addressed. Applicants should have greater motivation to fake the correct responses than would recent hires. Since content-valid items are verifiable, it is recommended that applicants in an operational situation be required to specify where they acquired their claimed experience.

PAPER SESSION

Focus on the Economics of Testing: Measuring Utility

Chair: Nancy Whitlock, South Carolina Interagency Merit System

An Evaluation of Alternative Methods of Estimating the Standard Deviation of Job Performance to Determine the Utility of a Test In A Fixed-Treatment Sequential Employee Selection Process

Robert S. Mayer, Wayne State University

The present research is designed to partially replicate the Schmidt et al., method for estimating SD_y by determining if the estimates of SD_y by supervisors correspond to estimates made by those being supervised and by estimates made by higher level personnel.

A second purpose of this investigation is to determine if an accounting estimate of SD_y corresponds to the estimates made by employees of the same organization.

A third purpose is to determine if the accounting estimate of individual employee output corresponds to an estimate of individual employee value made by their supervisors by a variant of the Schmidt et al. methodology for estimating SD_y. The variant of the Schmidt et al. methodology will entail supervisors being asked to estimate the value of specific individuals in their work group rather than estimating the value of hypothetical employees performing at the three performance levels.

Finally, this investigation is designed to continue the expansion of the body of research demonstrating the potential utility of a valid selection procedure.

The people involved in this study were 49 Head Tellers, 51 Branch/District Managers, and 92 Tellers employed by a large multi-branch bank. The results of this study indicated significant differences by organization level in the estimates of SD_y. The accounting estimate of SD_y corresponded only to the estimates made by higher level personnel. The accounting estimates of individual output did not correspond to the estimates of individual value made by supervisors. The estimated utility of the selection procedure was shown to vary considerably depending on which estimates were used for the parameters in the equation for assessing utility. These results indicated that the subjective method of estimating SD_y might not be the optimal solution for utility analysis. This investigation obtained further evidence that a valid selection device can offer an extremely cost effective technique for increasing the dollar gain to an organization. The objective criterion of job performance developed from the accounting data showed no relationship with the subjective criteria of job performance.

The second issue was the low correlation between the predictors available in this study and the objective performance measure. The reasons for the poor validities between the predictors and the objective criterion should be evaluated.

The final issue concerning the validity coefficient is using a coefficient of overall performance that represents the difference between the new procedure and the best a priori procedure. It might be that overall performance is not as relevant a criterion to the organization as some other criterion. If this is the case, the choice of which validity coefficient to use and determining whether alternative estimates of SD_y are appropriate to the validity coefficient that is used must also be considered.

This investigation obtained estimates of SD_y by two different methods. The methods were an approach developed by Schmidt et al. and a 'value-added' accounting approach developed specifically for this study. The results of this investigation indicated a lack of agreement between the two methods and between estimates of SD_y by different organizational levels with the Schmidt et al. method. These findings have rather disturbing implications for determining the utility of a selection procedure.

As there were alternative estimates of SD_y using the Schmidt et al. approach as a function of organizational levels, there are also alternative accounting methods for estimating SD_y. There are several other methods available for estimating the value of individuals in an organization that might offer potential as an alternative way of developing an accounting estimate of SD_y.

Finally, there was agreement in the evaluation of Teller performance evidenced by the positive relationships between self assessments and Head Teller assessments. The results of this study indicated that there were considerable differences between the accounting estimates of SD_y and the Schmidt et al. estimates. Investigations should continue on the construct validity of the standard deviation of job performance in dollar terms.

A final area for research estimating SD_y would be investigation of the situational specificity in the SD_y estimates. The generalizability of SD_y estimates to the same job in different organizations and/or job families could potentially have similar implications to the results found for the generalizability of validity of various predictor measures.

Personnel Selection in the Wake of Teal: Proposals for the Defense of Noncompensatory Systems?

Dennis Armstrong, Illinois Institute of Technology

Current procedures for estimating the economic impact of valid selection procedures have demonstrated substantial increases in workforce productivity.

Gains in workforce productivity due to the implementation of a valid selection procedure may run into billions of dollars, dependent upon the validity of the selection procedure, the size of the standard deviation of job performance in dollars, SD_y, the size of the workforce, and the cutoff score used for selection. These gains may be additionally increased if one considers the tenure of thos selected (Schmidt et al., 1979).

The purpose of this study was to compare the utilities of two selection systems. The first selection system is a special case of the non-compensatory system which has been referred to as the "knock-out" (KO) system. The selection devices in a KO system are applied sequentially, instead of simultaneously. Thus, applicants who fail to meet the minimum cut-off at the first stage of a selection process are rejected, and are not given the opportunity to participate in the subsequent stages of selection.

This KO system was compared to a compensatory (CMP) selection system. The results indicated that the utilities per selectee for KO systems did not compare favorably with the utilities for a comparable CMP system. At best, the utilities for a KO system begin to approach the utility for a CMP system, dependent upon the validity of the test and selection ratio at stage one.

The analysis of alternative selection models indicated that implementation of KO systems can result in increases in total workforce productivity. The use of "optimal" selection models, given a KO system, resulted in significant gains in workforce productivity in comparison to random selection procedures. The gains due to the use of the best KO selection model, however, are not expected to approach gains due to use of a comparable compensatory selection procedure.

Test Utility for Mechanical Jobs

Murray J. Mack and Richard H. McKillip, Psychological Services, Inc.

The purpose of this paper is to describe the results of two criterion-related validity studies with particular emphasis on test utility. As such the number of occupations along the job spectrum for which utility analyses have been conducted is expanded. In performing these analyses it is possible to compare and contrast the utility of selection procedures as currently used administratively by the organization versus the utility that could be attained through relatively minor modifications in test use.

Criterion-Related Validation. The two criterion-related studies took place at a major public utility company located on the eastern seaboard. The job progressions covered by these studies were mechanically oriented in nature and were covered generically under the Dictionary of Occupation Titles, job code 953.884, Gas Meter Installer. The two job progressions were Service Specialist and Street Mechanic.

The criterion measures used in these studies consisted of two multiple-choice problem situation tests. These measures differed from the usual tests of job knowledge in that special care was taken to ensure that the majority of items measured the ability to troubleshoot problems encountered on the job.

Criterion-related validities (adjusted for unreliability in the criterion and restriction of range in the predictor) were .68 and .82, respectively, for Service Specialists and Street Mechanics ($p<01$). Uncorrected coefficients were .46 and .76.

Utility Considerations. The utility estimates were derived using actual data collected on the validation research and applicant samples. Applicant test data were analyzed for an 18-month period prior to the study. There were 360 applicants taking the tests over this period. Fifty were selected for one occupation, and thirty-seven for the other.

The basic equation for determining test utility is as follows:

$$\Delta \bar{u}/\text{selectee} = r_{xy} SD_y \bar{Z}_x - c/p$$

where

$\Delta \bar{u}$/selectee = the average gain in productivity in dollars per year per person selected as a result of using the valid tests;

r_{xy} = the validity of the new selection procedure; the correlation between scores on the procedure and job performance for a group of incumbents;

SD_y = the standard deviation of job performance among incumbents;

$\bar{Z}_x$ = the average standard score on the selection procedure for those selected;

c = cost of testing per applicant; and

p = the proportion of applicants selected (the selection ratio)

According to the research by Schmidt-Hunter and their associates, SD_y can be appropriately but conservatively estimated as approximately 40% of average annual salary.

Service Specialist. The estimated productivity gain from using the selection test battery for each Service Specialist in this career path selected over the last eighteen months is $5007.68. The average gain in productivity for the entire group of these 37 hired, is adjusted for a twelve month period, $125,192.00. These gains in workforce productivity are gains for one year's worth of job performance. Multiplying the average gain in productivity for each selectee by a tenure factor will provide an estimate of workforce productivity over the course of a career.

Street Mechanic. The average gain per person selected for this occupation over the past eighteen months is $295.27. The average gain per year for the group of 50 adjusted for twelve month period is $9,743.91.

Discussion of Utility. These results are quite compelling. The criterion-related validities differ for the two occupations (.82 compared to .68). In addition, the company has been hiring more of one group than the other as noted by the selection ratios (.14 to .10 respectively). Yet the company is losing thousands of dollars on a per person basis ($5,007.68 compared to $295.27). What is causing this difference in test utility?

The answer lies in the administrative use of the selection test battery. Close inspection of the average group test scores reveal over a twenty point discrepancy. The company is making effective use of the test battery in hiring for one occupation; however, the hiring practices for the other group are essentially random. The average test score for this "random" group is only one point greater than the average test score from the total applicant pool. The result is that the company is taking some advantage of the potential gain from the use of valid testing for one group but not the other. One way that the company can increase the job productivity of its workforce is by making systematic selection decisions according to either previously established or newly set test cutoffs.

Comparison of Benefit Variance Estimation Procedures for Determining Utility

Catharine W. Burt, U.S. Bureau of the Census

It is a widely accepted view that expressing the validity of a selection procedure in terms of economic gain to an organization is a valuable interpretation of the worth of the procedure. This economic value is known as the utility of the procedure in selecting qualified personnel. A common definitional formula for calculating test utility per selectee is the product of the validity coefficient (r_{xy}), the average standard score on the test of those hired ($\bar{Z}_{xs}$) and the standard deviation of job performance in dollars (SD_y). The product of r_{xy} and $\bar{Z}_{xs}$ is the mean standard score on the dollar criterion of those selected using the test, $\bar{Z}_y$. Therefore, utility per selectee would be the mean Z-score on the criterion of those selected times the standard deviation of the criterion in dollars. It should be clear that because of its multiplicative effect, SD_y has a critical role in determining the utility of a selection procedure.

Many studies have shown the difficulty in estimating the standard deviation of employee benefit to an organization. Charles Schultz (1980) and Frank Schmidt (1979) have presented two techniques for estimating cost variability for use in the formula for determining test utility. Schultz's technique was based on the premises that the most productive worker accomplishes about three times as much work as the least productive worker and that the value of the average worker was the prevailing pay for the job. He then built the ratio of 3:1 into his calculation of range from most to least productive worker. Schmidt presented another estimation technique in which he asked experts the value of productivity at the 15th, 50th and 85th percentile. By averaging these judgments, an estimate of one standard deviation can be made if an assumption of a normal distribution is made.

In determining the variability of benefit of work accomplished by census enumerators, the author thought that the assumptions made in these two estimation techniques were not necessarily valid. A third technique was examined in which the assumption was made that the number of standard deviations between the highest and lowest scoring employee on the performance measure was equivalent to the number of standard deviations between the most and least cost-effective employee. Cost-accounting techniques were used to determine the benefit of these two employees. The range between them was divided by the number of standard deviations between the highest and lowest performer on the performance measure in order to find the standard deviation of benefit.

These three estimation techniques were compared with an actual cost-accounting technique for determining the standard deviation of employee benefit in terms of dollars necessary for use in calculating test utility. The techniques were applied to validity data collected during the 1980 census on a sample of enumerators. The results showed differences in the estimated standard deviations with the largest difference provided by Schultz's technique. Schmidt's technique was found to be reasonable only if the lower limit of a 95 percent confidence interval around SD_y was used. A third technique proposed in this study seemed to be very reasonable in estimating SD_y. The use of this technique might be found to be more reasonable if easier methods could be utilized to determine the benefit of the most and least effective performers.

PAPER SESSION

Training and Experience Evaluations and Other Forms of Self-Assessment

Chair: Carol Granfield, City of Fairfax
Discussant: Stephen E. Bemis, Psychological Services, Inc.

Assessment for the Selection and Recruitment of Public Health Nurses

Lourdes Lacomba, S.C. State Personnel Division Merit System

Public health nurses in South Carolina are assigned to diverse program areas such as Maternal and Child Care, Immunization, or Home Health Services. As a result, the emphasis given to particular duty areas is influenced by the program in which a duty is performed. In the present study, a structured job analysis instrument was developed for public health nursing in order to adequately survey various nursing job dimensions. By highlighting the similarities and differences in task performance of incumbents within a particular job group, it was felt that better judgments could be made in finding candidates who best match the needs of the job group.

A task base structured job analysis tool, the Public Health Nurse Questionnaire, (PHNQ), was developed to capture any context specific differences as well as to generalize across nursing duty areas. Out of a total population of 1,011 incumbents, a stratified sample of 630 nurses were asked to complete the PHNQ. Study participants were distributed across 34 locations, 16 programs, and 8 positions. The 472 questionnaires that were returned, or 75% of the forms, succeeded in covering all the functional areas of nursing.

Examples of task items included in the PHNQ are:

> Observes a patient's general appearance including his/her hygiene, awareness, mood, color, gait, activity level, signs of distress and speech for evidence of symptomatology.
>
> Develops care plan in collaboration with patient, significant others, or members of the health care team which describes patients' problems. care objectives and strategies for resolving patient needs.

During the PHNQ rating exercise, nurses were instructed to rate each task on a 3-point importance and entry-level scale, and on an 8-point frequency of performance rating. The importance and entry level scales were included in the rating exercise so that a set of performance differentiating tasks that can be performed upon entry could be identified.

To establish the job dimensions, the ratings on the frequency of performance scale were submitted to a principal factor analysis with the squared multiple correlation used as the estimate of communality. The factor structure was rotated by applying the VARIMAX criterion. Grouping incumbents on the basis of their factor profile distance values was achieved by Ward's hierarchial cluster analysis.

Reliability of the PHNQ. An intraclass reliability estimate of r =.96 resulted on the basis of all study participants' responses to the frequency of performance scale.

Factor Analysis. The 8-factor solution which accounted for 74% of the variance was selected as producing the most intuitively clear summary of the task items. The resulting job dimensions were defined as 1) Health Status and Planning, 2) Physical Assessment, 3) Nursing Service Delivery, 4) Program/Services Management, 5) Lab Processing, 6) Specialized Procedures, 7) Program Compliance and 8) Prenatal Examination Procedures.

Job Profile Comparisons. The standardized dimension means for several clusters were plotted to illustrate the similarities and differences of task performance across the eight job dimensions and how these plots may be interpreted. The mean profile for Cluster 7 is comprised of mainly nurses who work in programs found in mental health centers. This setting may explain their low means on the Physical Appraisal and Lab dimensions which cover health clinic type activities. The means for Cluster 9 contain mainly Home Health nurses who provide care to medical-surgical patients in their home. The high mean on nursing services delivery reflects the emphasis that these nurses give to this dimension. Similar analyses for other settings/programs were explained.

Discussion. The program and position level attributes of each cluster helped to confirm that task performance was likely to be a function of a nurse's job assignment. Reporting job experience in the profile format would enable supervisors to quickly identify a candidate whose experience most closely resembles the work that is performed by the incumbents in a particular unit. The job dimension mean scores, which serve as standardized job content descriptions, provide an objective basis for assessing the degree of relevance that an individual's experience has in various nursing contexts.

A Comparative Study of the Behavioral Consistency and Wholistic Judgment Methods of Job Applicant Training and Work Experience Evaluation

Ronald A. Ash, School of Business, University of Kansas

The evaluation of job applicant training and work experience (T & E evaluation) constitutes one of the most widely used personnel screening and selection procedures. There are at least six different methods of T&E evaluation:

1) *The Point Method*. The traditional point method is clearly the most prevalent one used by public sector jurisdictions. It consists essentially of a mechanical formula set out in a formal schedule. Points are credited for the number of months or years of different kinds of relevant training, education, and experience.

2) *The Grouping Method*. In the grouping method applicants are usually divided into a samll number of groups (from two to seven) on the basis of the *simultaneous* consideration of training and experience. Applicants assigned to each respective group are assigned the same score.

3) The Task-Based Method. This approach relies on the premise that greater validity can be achieved by obtaining detailed information on specific tasks which an applicant has performed in the past, regardless of the job in which the tasks were performed. The task-based method is operationalized by means of a supplemental application form consisting of a task inventory.

4) The Knowledge, Skill, and Ability (KSA)-Based Method. The most common KSA-based approach to T & E evaluation is the "applicant self-report checklist."

5) The Behavioral Consistency Method. The goal of the behavioral consistency method is to rank-order applicants on the basis of the kind of achievement behaviors that are required for superior performance on the job in question. The KSAs are combined into major achievement dimensions, usually from five to seven. For each major achievement dimension, applicants are encouraged to describe in detail at least two past achievements which best demonstrate their capabilities.

6) The Wholistic Judgment Method. The typical method for evaluating information on job applications or resumes involves some combination of the employment manager or other senior personnel worker and/or the hiring line manager to review the applicants' statements of physical, educational, and experience qualifications in relation to the requirements of currently available jobs.

This paper presents results from an exploratory study comparing a behavioral consistency method with a wholistic judgment method in terms of rate-rerate reliability, interrater reliability, T & E evaluator scoring time, and applicant completion time. Applicant behavioral consistency scores are also compared with applicant ranks from the wholistic judgment method, and both of these are compared to applicant scores on the Wonderlic Personnel Test.

The job applicants were 47 college students being considered for the job of Planner in the Product Management Division of a large corporation. Applicants were asked to complete the corporation's standard professional employment application form, a job-related achievements application supplement, and a brief form indicating the amount of time they spent completing each. One graduate student and two upper division undergraduate students served as T & E evaluators. The T & E evaluators independently ranked the 47 job applicants in terms of their respective suitability for the job of Planner using only the information contained in the standard application form. This was the operationalization of the wholistic judgment method. This was repeated a second time one week later.

Subsequently, the T & E evaluators independently scored the job-related achievements application supplements using the benchmark achievement rating scales. This was the operationalization of the behavioral consistency method.

Rate-Rerate Reliability. The rank-order correlations obtained for the wholistic judgment method were .70, .77, and .82, yielding an average rate-rerate rank order correlation of .77. The correlations obtained for the behavioral consistency method were .95, .94, and .94, yielding an average rate-rerate reliability coefficient of .945. The behavioral consistency method resulted in higher rate-rerate reliability than did the wholistic judgment method. The average wholistic judgment rank order correlation of .825 is not significantly different from the behavioral consistency average interrater reliability coefficient of .84. Thus, overall, the two methods do not appear to differ in terms of interrater reliability.

Comparison of Behavioral Consistency Scores with other Types of T & E Scores. For the most part, the correlations of behavioral consistency scores with "credentialistic" T & E methods--a term used by Schmidt et al. to describe KSA-based and grouping methods, and which might be applied to traditional point methods as well--are essentially zero. However, in every instance where behavioral consistency and task-based scores are compared, the correlations are moderate and positive. It appears that behavioral consistency and task-based methods are measuring things which overlap in part.

T & E Methods and General Mental Ability. The behavioral consistency method puts greater demands on applicant verbal, writing, and communication abilities than do typical application forms. Given the high relationship between verbal ability and general mental ability, one might expect behavioral consistency scores to correlate more highly with the Wonderlic scores than would wholistic judgment scores. In the current study both correlations obtained are essentially zero.

Conclusion. The primary weakness of this study is the absence of empirical validity data. Reliability of both the T & E evaluation methods used in the present study appears adequate, and both methods apparently yield information nonredundant with that obtained from general mental ability tests.

Reexamination of Self-Assessment for Employee Selection

Benjamin Ocasio, U.S. Postal Service

This study investigated the predictive validity of self-assessment of abilities for employee selection using job applicants. The specific hypothesis tested was that self-assessment of custodial abilities would not be significantly correlated with supervisor's ratings of corresponding abilities.

A task analysis was conducted for the position of Custodial Laborer in 189 offices throughout the United States, using a sample of 379 employees. The employees rated 32 custodial tasks on time spent and importance. The primary purpose of the job analysis was to design two job-related questionnaires; a four point self-assessment questionnaire and a five point supervisor-performance evaluation questionnaire. Both questionnaires consisted of the 18 activities described in the job analysis as critical and important.

Also a question on the employee's demonstrated overall performance was used as a second measure of job performance. Sixty-seven custodian applicants from Houston, Texas, asked to rate their current ability on each of the 18 items. Three months after the applicants were hired, their immediate supervisors rated them on the supervisor rating questionnaire. The supervisors completed the questionnaire independently and without knowledge of the employee's ratings.

Some interesting differences were noted between the ratings made by supervisors of the employee performance and the ratings completed by the employees themselves. No employees were rated as poor. It was hypothesized that the applicants would afford themselves higher ratings than those they received from their supervisors. Comparisons of the average mean ratings rendered by supervisor - and self-ratings show that most of the self-ratings were lenient. The average mean rating for supervisor-ratings was .22 scale units below the expected mean of 2.5. The average mean rating for self-ratings was .47 scale units above the expected mean of 2.5. It is evident that the self-ratings are more lenient than the supervisor-ratings. As had been hypothesized, the applicants have a relatively high opinion of themselves.

In order to examine the validity of self-assessments, supervisors evaluated 67 employees on the 18 performance activities and one measure of overall job performance. An overall nonsignificant correlation was found between self-assessments and supervisor ratings ($r=-.07$, $p>.05$). Of the 18 correlations between self- and supervisor-ratings, only six were positive correlations, all are nonsignificant.

PAPER SESSION

Exploring Issues Related to Fairness, Adverse Impact, and Test Bias

Chair: Stephen E. Bemis, Psychological Services, Inc.
Discussant: Bruce W. Davey, State of Connecticut

The Effect of Preselection on Adverse Impact Determination

Walter G. Mann, Office of Personnel Research & Development

The number of qualified applicants decreases at each hurdle of a multiple-hurdle selection process. This paper will refer to the selection that precedes a particular hurdle as "preselection." Although there is no literature on preselection, it seemed reasonable that the greater the preselection, the less the adverse impact.

7443 individuals took the standard GS-7 examination, a written test, and returned the race, sex and national origin (RSN) questionnaire. Comparison of the RSN percentages shows that most RSN groups had representation ratios exceeding 1.00, with only whites and females underrepresented. For example, 6.1% of all applicants in the sample were Hispanic. The comparable figure in the civilian workforce was 4.2%. When the former is divided by the latter the result is 1.45.

The number of applicants in various RSN groups was counted. Then the number of selected applicants in each RSN category was counted. A selection rate for each RSN category was computed (number selected divided by number of applicants). Because whites and males had the highest selection rates, selection rates of minorities were compared with that for whites, and the selection rate for females was compared to that for males. Evidence of adverse impact was of two types: size (4/5's rule) and statistical significance (.05 level). The results clearly demonstrate the written test has considerable adverse impact for minorities.

The adverse impact analysis of specialized experience revealed no adverse impact against minorities; minorities were just about as likely to receive bonus points for specialized experience as were whites.

A key difference in the analyses of the written test and specialized experience is the base group. The written test statistics were based on all applicants who took the test. This is shown in tables below.

Adverse Impact Analysis of the Written Test
When Used as a Minimum Qualifier (N=7354)

	Minorities	Whites
Number of Applicants	2432	4922
Number passing	501	2663
Passing Rate	.206*	.541
Adverse Impact Ratio	.38[a]	-

*Significantly different (.05 level) from passing rate for whites.
[a]Adverse impact according to the 4/5's rule.

Conclusions. It is clear that the context within which a selection device is used has an overwhelming effect on the estimation of adverse impact. It is important to know the particular function of the device, e.g., screening vs. ranking. It is also important to know the amount of pre-screening on the sample that has taken place. This pre-screening can result from self-selection on the part of applicants, as well as from other selection devices.

Minimizing Adverse Impact While Maintaining a Merit System

Deborah P. Ashton, California School of Professional Psychology At Berkeley and Joel P. Wiesen, Bureau of Planning and Research, Common-Wealth of Massachusetts

In developing the November, 1981 entry-level police services examination for the Commonwealth of Massachusetts, great care was taken to produce an examination which was valid and at the same time minimized adverse impact. An innovative approach was used to reduce adverse impact while maintaining the integrity and utility of the test instrument. This IPMAAC paper discussed this innovative technique and suggests other techniques which can be implemented to reduce adverse impact while maintaining a merit system of personnel selection.

The November, 1981 examination tapped the same areas as the October, 1975 police service examination (developed by IPMA). Since the 1975 police services examination was found to be highly predictive (unadjusted correlation coefficient, (r of .49, $p<.001$, adjusted r of .71) of police academy training grades. Since the November, 1981 police services examination is based on the 1975 examination in content, and is strongly correlated with it (raw score correlation = .71), the 1981 examination has predictive and content validity. Also, the 1981 examination is based on the construct and content validity study performed in the development of the 1979 police services examination. This study confirmed the content validity of the examination, and explored the construct area. Therefore we believe that use of the 1981 examination supports a merit system of personnel selection.

In developing items for the 1981 examination, precautions were taken to safeguard against cultural bias. First, a cross section of people (by ethnic background, race and sex), placed in various situations, was represented in the items developed. Second, members of minority groups were not presented in stereotypic roles. Third, racial and ethnic groups were also socially stratified in a multitude of situations, in the test items which required such narrative descriptions. Cultural review guidelines issued by the Commonwealth and other sources were used to guide the examiner staff in the test construction stage. Importantly, a report commissioned by the Commonwealth of Massachusetts, "Guidelines For The Identification and Elimination of Test Bias in Standardized Tests" (Flaugher, Nieves, Slaughter, Wiesin and Woodford. 1979, was used as an aid in designing test items; particularly the guideline's section "Modes of Offensiveness in Test Items." Items which are offensive are items which are particularly offensive to minority group members. The specific ways in which an item may be offensive was discussed in some detail.

In addition, an innovation was used in the selection of items in order to reduce adverse impact. The construction of the examination incorporated an effort to statistically reduce adverse impact by selecting test items which had been pretested and which yielded at least an 80% minority to non-minority success ratio.

At the recommended passing point of 70, the achieved ratio of minority to non-minority passing rates was .69. This is the highest such ratio for all police service examinations in Massachusetts since 1975. Previous ratios ranged from .48 to .58.

Finally, the possibility of reducing adverse impact further was discussed. It was suggested that if criterion data is available, the subtest scores may be individually inspected for adverse impact, validity and intercorrelation. It may be that some subtests are highly intercorrelated and that the weight of subtests with high adverse impact could be reduced without greatly affecting validity. Although the test would be weighted to reduce adverse impact, it may be possible to find weightings which maintain the validity, based on empirical methods. The underlying assumption to this proposed technique is that the validity coefficient is robust enough to allow such reweighting of subsections.

Bias In Content-Valid Tests Revisited

John G. Veres, III, Wiley R. Boyles, Cecilia H. Champion,
Auburn University at Montgomery

Bias in job content domain occurs when job content differs for different subgroups of incumbents. When two different jobs are treated as a single job for the purposes of test development, any sampling of one job domain is unfair to individuals applying for the other. A prior study by Wiley Boyles, Chester Palmer and myself found significant differences between job analysis ratings produced by black subject matter experts and those produced by white job experts on the same clerical jobs. Black applicants scored highest vs. white applicants on those areas of the selection test that black incumbents had rated lower than white incumbents on the job analysis.

A further study to investigate this matter was conducted. Forty of the 126 state government clerical workers that participated were black clerks and 86 were white. Subjects were drawn in groups, each group working in the same office and performing essentially equivalent job duties. All subjects were designated as performing substantially equivalent duties by the state's classification plan. Each subject rated his/her own job and the job of one or two coworkers.

The job analysis task statements for rating were adapted from a validation study of a clerical written examination administered by the same state. The broad responsibilities assessed in the job analysis were preparation of typed documents, use of filing systems, and communications. Tasks provide operational definitions of the responsibilities. In addition, three tasks were included that were not performed by clerical workers. This follows the advice of Stutzman and Green, who used low-frequency tasks as a means of assessing rater accuracy.

Five rating scales were provided for evaluating each task required by the job under review: frequency of occurrence, time required to perform the responsibility per occurrence, criticality to successful job performance, extent necessary upon entry to the job and ability to discriminate among varying levels of performance.

Results. Profile analyses were used to test for racial differences in SME descriptions of job content. Race was the independent variable. Each of the three responsibilities (typing, filing and communicating) used to describe the job was examined separately. Incumbents displayed no significant differences in profile shape. That is, black SMEs and white SMEs reported performing essentially the same duties.

No definitive conclusion can be reached concerning the relationship between racial bias and incumbents' job analysis rating behavior. The two pieces of empirical research extant indicate that racial differences in job analysis ratings can occur. Future research is required before any statement regarding fairness can be made with any degree of confidence.

PAPER SESSION

Assessing Managerial Skills

Chair: Terry S. McKinney, City of Phoenix

Development of Promotion Evaluation Procedures For A Centralized Referral System

Garth Wall, U.S. Army Civilian Personnel Center and
Jay A. Gandy, U.S. Office of Personnel Management

The U.S. Office of Personnel Management's Office of Personnel Research and Development and the U.S. Army Civilian Personnel Center (CIVPERCEN) recently conducted a joint agency effort to develop and test methods and procedures to improve Army's centralized candidate evaluation and referral system. As a pilot effort, this project focused on the civilian personnel administration (CPA) career field.

A job analysis approach was needed which would permit valid assessment of candidates both within and across specializations. The CPA career field is made up of six specializations, including general personnel management, staffing, classification and pay, employee relations, labor relations, and employee development. We selected the work activities inventory-survey job analysis approach as a means of comparing jobs at the behavioral level. There were eleven detailed steps involved in the survey development.

Analysis Procedures for Work Activities/Survey Results. Results were processed using the Comprehensive Occupational Data Analysis Programs (CODAP) (Christal, 1974). We used principally two types of analysis: group comparisons based on background variables and cluster analysis. The group comparisons entailed comparing composite job descriptions of groups identified on the basis of variables such as job type, grade level, or level of supervision. For example, we compared three levels of supervision in terms of most frequently performed work activities and the relative importance of the work activities to each level of supervision.

SME Panel Procedures. Each panel carried out a four-step process under the leadership of a member of the project team. The process was designed to (1) determine which work activities differentiate superior performance; (2) determine knowledges and abilities which differentiate superior performance for the specific work activities identified in step 1; (3) determine knowledges and abilities which differentiate superior overall job performance; and (4) recommend refinements in knowledge and ability definitions and reconcile differences among panels.

Analysis of Data from SME Panels. The reliability of panel ratings was estimated by applying analysis of variance procedures for average ratings as recommended by Ebel. While the reliability values are not uniformly high, we believe they are satisfactory in view of the fact that the knowledges and abilities, and also the work activities rated, had been selectively narrowed prior to the ratings. We found that, except for one ability related to supervision, the abilities considered measurable were applicable to all of the jobs regardless of specialization.

Development of Evaluation and Referral Procedures. Concurrently with the job analysis effort, a general review was undertaken of the state of the art of candidate evaluation techniques and the referral process. This was described in detail.

Design and Implementation of the New York State Management Skills Inventory

Robyn Katz, Vincent Perfetto, James Moon, New York State
Department of Civil Service

Project Background. In December of 1981, the New York State Department of Civil Service began exploring the feasibility of constructing a computerized skills inventory that would contain detailed information on the State's top level program managers. The first step in developing the skills inventory project proposal was to conduct a needs survey of potential system users. One major concern was voiced by a number of respondents, regarding the questionable reliability of information in such a system, particularly if program participants would be reporting on their own qualifications.

System Design. Design activities focused on three major areas: 1) identifying the data elements to be included in the file and designing the data collection instrument and file record layouts, 2) developing a controlled vocabulary to describe both participants' backgrounds and position requirements, so that matches could be effected by the system, and 3) designing the search and retrieval capabilities of the system.

The system was designed to: 1) perform both broad, general searches and narrow, specific searches, depending on user need. For example, we wanted to be able to identify everyone in the file with a Master's degree and administrative experience, but we also wanted to be able to identify anyone in the file with a Master's degree in industrial psychology and three years of middle management experience in personnel selection. 2) identify individuals who met certain minimum requirements, and to select from among that group individuals with additional, desirable qualifications. 3) closely parallel the structure of the minimum qualifications for State titles.

System Implementation. The major system implementation activities included: 1) data collection, 2) data preparation and entry, 3) completion and testing of system programming, 4) announcement and marketing of the system, 5) development of procedures for operating and maintaining the system.

We began by collecting information from a subset of top level State managers, numbering approximately 2,300. The initial group invited to participate in the Management Skills Inventory were all career, competitive class employees, earning in excess of $36,000 a year as of June of 1982 (the cutoff has since been increased to $39,000, based on a later 9% raise).

The initial response to the program was quite encouraging. We received more than 500 responses during the first month, with returned questionnaires outnumbering declinations two to one. However, the number of responses declined sharply in early September, and we began receiving less than 50 questionnaires a week. To our surprise, it appeared that we would fall short of our initial estimate of 1,200 or more program participants.

Current Status. The Management Skills Inventory was operational by the October 1982 deadline, with approximately 500 individual records in the file. As of now, we have 650 individuals in the file, and are receiving and processing additional questionnaires as the result of one final follow-up with non-participants and our first update identifying new individuals in the top management ranks. The agencies for which we have demonstrated the system were uniformly impressed and have expressed considerable interest in having the program expanded to cover all of their management employees. However, use of the Skills Inventory is not yet widespread.

PAPER SESSION

Alternatives and Traditional Selection Procedures

Chair: Jurutha Brown, City of Los Angeles

The Development and Evaluation of Three Traditional and Three Alternative (Non-Traditional) Selection Procedures

Leaetta M. Hough, Marvin D. Dunnette, and Margaret A. Keyes,
Personnel Decisions Research Institute, Minneapolis

The Uniform Guidelines on Employee Selection Procedures assert, among other things, that when a valid selection procedure shows adverse impact, the employers should develop alternative, equally valid, selection procedures--procedures having less adverse impact. A consent decree entered into by a federal regulatory agency provided an opportunity to develop and evaluate selection and promotion procedures for attorneys that included both traditional predictor instruments and non-traditional procedures (alternative procedures). The empirical evaluation of both the traditional and alternative inventories consisted of a criterion-related, concurrent validation study.

Job Analyses. Job analyses guided the development of the predictor instruments. Information about the attorney position at the agency was gathered using three procedures. A structured task checklist was developed, pre-tested, and then administered via mail.

A behavior-oriented job analysis using Flanagan's Critical Incident method was also done. In addition, interviews with 25 key agency officials were conducted. These officials described the personal characteristics important for success as the agency and identified prior experiences likely to be indicators or predictors of success at the agency.

Description of Predictor Instruments. The traditional predictor instruments were based on information obtained during the interviews with high level agency officials. They consisted of 1) a Background Information Inventory, which were questions of objective, verifiable biodata type items. The questions asked about schools attended, college grades, law school grades, honors, publications, speeches, club memberships, etc. 2) Interest and Opinion Inventory containing questions which were a blend between biodata type items and personality items. They represent our effort at tapping the constructs that agency officials said were important for success at the agency, 3) Self Description Inventory which is a standard personality inventory developed by Edwin E. Ghiselli. The purpose of this inventory was to measure, very quickly, several of the personal characteristics mentioned by agency officials as important for success at the agency.

The three non-traditional or "alternative" predictor instruments were also based directly on the job analysis results--the task job analysis and behavioral job analysis. They consisted of 1) a Situational Judgment

Inventory. The Situational Judgment Inventory was based on critical incidents that were gathered during the behavioral job analyses. Using actual incidents, realistic job situations and possible actions were written. For each situation, the attorneys were asked to indicate which one was the best action and which was the worst, 2) a Task Importance Inventory based on the results of the structured task checklist job analysis. The attorneys were asked to rank order 14 tasks in each of three sets in terms of their importance for success at the agency, 3) Accomplishment Record Inventory which was based on the results of the critical incident method of job analysis. Eleven behaviorally defined dimensions emerged from the critical incident data. We selected eight of these dimensions for the Accomplishment Record Inventory. Attorneys were asked to describe major past accomplishments illustrative of their competence in each of the eight job areas. We gathered approximately 2600 accomplishments that needed to be scored. We used a two-step procedure to develop scoring guidelines.

Development of Criterion Measures. Task-oriented rating scales were developed using the task/activity checklist job analysis results. During the analysis of the task checklist, salience scores (the arithmetic sum of time spent and importance ratings multipled by percent saying the task was part of the job) were computed for each task. For each factor of the job identified by factor analyzing the time spent responses, highly salient tasks were identified. These highly salient tasks formed the task-oriented job performance rating scales. In addition, we used Flanagan's critical incident method and Smith and Kendall's "retranslation" approach to define the behavioral dimensions of the job and develop behaviorally-anchored performance scales.

Data Collection. The six predictor inventories were administered under a variety of conditions. Some attorneys attended group sessions where they completed the battery, some received group instruction but completed the battery as time permitted, others received the battery in the mail. Three hundred twenty-nine attorneys (67%) returned completed or partially completed predictor batteries.

The performance appraisal rating forms were distributed to supervisors via the mail. Performance ratings were obtained for 390 attorneys (79%). Both predictor information and performance ratings were available for 267 of the attorneys (54%).

Evaluation Results. We formed a composite for the criterion and for each of the six predictor inventories. The overall criterion composite was formed by equally weighting the mean overall task rating and the mean overall behavioral rating. The reliability of the mean overall task rating was .83 for attorneys with two raters (N=150); the reliability of the mean overall behavioral rating was .70 for attorneys with two raters (N=153).

Traditional Predictor Inventories.

1) Background Information Inventory. Recall that we rationally formed scales within the Background Information Inventory. We correlated these rationally scored scales with the overall criterion and used these correlations to weight the scales. Thus, we validity-weighted the rationally formed scales to obtain a composite. Though the capitalizing on sample specific variance was less than if we had used multiple regression analyses, cross validation was still warranted. Unfortunately, the available shrinkage formulas overestimated the shrinkage in this case; therefore, we chose to estimate the cross validity with a Monte Carlo technique. The estimated cross validity of the Background Information Inventory against the overall criterion composite was .37.

2) Interest and Opinion Inventory. The composite for the Interest and Opinion Inventory was obtained in the same way as the Background Information Inventory composite. The estimated cross-validity against the overall criterion composite was .25.

3) Self Description Inventory. Validity-weighting the already existing scales to form a composite and correlating that composite against the criterion composite yielded a correlation that was essentially zero; therefore, we combined the scales using multiple regression analyses. Using the shrinkage formula suggested by Cattin, the cross validity was .05. Of the traditional inventories, the personality inventory did least well and the biodata inventory did the best.

Non-Traditional Predictor Inventories.

1) Situational Judgment Inventory. We intended to validity weight the items; unfortunately the medium validity of the items was only .04. We concluded that the content-oriented strategy used in developing the situational judgment items was unsuccessful. It is possible that a different scoring system would have yielded more promising results.

2) Task Importance Inventory. A composite score was derived by summing the scores on the three parts of the inventory. Unfortunately, the composite correlated only .02 with overall performance. Again, we concluded that the content-oriented strategy used in developing this alternative inventory was unsuccessful.

3) Accomplishment Record Inventory. The composite for the Accomplishment Record Inventory was simply the mean overall evaluation on each attorney's accomplishments. In analyzing the Accomplishment Record scores it became apparent that attorneys with many years of professional experience tended to obrain higher mean scores than those who had fewer years of experience. Obviously, attorneys with considerable experience had had more opportunity to do things of consequence than attorneys with more limited experience. To take this into account an "adjusted" accomplishment was used.

The correlation between the adjusted overall Accomplishment Record scores and the overall criterion was .25. Since mathematical optimization strategies were not used, estimating the cross validity was not necessary.

Only one of the alternative or nontraditional inventories demonstrated criterion-related validity.

Intercorrelations of Predictor Instruments. The intercorrelations between the three useful instruments, the Background Information Inventory, Interest and Opinion Inventory, and the Accomplishment Record Inventory, were low to moderate.

The multiple correlation of the Background Information Inventory composite, the Interest and Opinion Inventory composite, and the Accomplishment Record Inventory composite against the overall criterion composite was .49. This correlation was not corrected for unreliability in the criterion or restriction in range.

Summary. Two of the three "alternative" procedures developed for this study failed to show useful criterion related validities, whereas two of the three traditional procedures did produce useful empirical validities. The Accomplishment Record Inventory was the only "alternative" or non-traditional inventory that showed concurrent validity with job performance. One of its advantages, in comparison with, for example, the Background Information Inventory, is that it is a dynamic rather than a static predictor. That is, persons can improve their scores over a period of time by accomplishing more impressive achievements, whereas, an individual's score on a biodata inventory remains the same for life. Another advantage is that the Accomplishment Record is minimally correlated with more traditional predictors such as biodata inventories and aptitude tests and yet correlates with on-the-job performance.

Though we do not recommend using the Accomplishment Record Inventory alone (overall validity is greater when used in concert with more traditional inventories), the kinds of information tapped by Accomplishment Record Inventory appear to provide a unique contribution to a selection system.

The Development of Some Innovative Predictors for Entry-Level Selection

Ronald J. Karren, University of Massachusetts

With the adoption of the Uniform Guidelines on Employee Selection Procedures, an important concern of those in the selection field was whether alternative predictors of job performance could be utilized that would simultaneously maximize validity and utility and minimize adverse impact.

At the Immigration and Naturalization Service of the U.S. Department of Justice, there was a concern over the oral interview phase of their selection process for entry-level Border Patrol Agents. This interview was unstructured and unvalidated, and it was decided that this part of the selection process should be reevaluated and modified.

A job analysis was performed, and subject matter experts (supervisors) developed and rated both critical work activities and important knowledges, skills, abilities, and other characteristics (KSAOs). Considering both feasibility and assessment of the more important KSAOs, the following three predictors were chosen for development: an expected-motivation questionnaire, a miniature training and evaluation test, and an interview.

Expected-Motivation Questionnaire. Most personnel researchers assume that job performance is determined primarily by a person's work motivation and ability. Nevertheless, tests and selection devices in use today generally focus on the assessment of job-related knowledges, skills, and abilities, and little attention is given toward predicting the applicant's job-related motivation. The primary purpose of the expected-motivation questionnaire is to produce an estimate of the applicant's job-related motivation. It is derived from congruence theory, which postulates that work motivation is determined by the fit between the various characteristics of the job and the individual's work related motives and interests. Past research on the congruence approach has demonstrated both validity and cost effectiveness.

The first step in the development of the expected-motivation questionnaire (EMQ) was compiling a large pool of job characteristic statements. Characteristics that appeared motivationally relevant and specific to the Border Patrol job were considered eligible for the pool. The final pool contained 118 characteristics.

These characteristics were then rated by 75 Border Patrol Agents for both personal desirability and appropriateness or presence in the Border Patrol entry-level job. On one questionnaire, incumbents were asked to rate each characteristic on a 9-point scale ranging from "Extremely Undesirable" to "Extremely Desirable." In the second questionnaire, each incumbent rated each characteristic on presence in or similarity to the Border Patrol job on a 4-point scale which asked "How similar is the job described in each item to this job?"

The resulting data was analyzed to obtain the mean desirability rating and the mean similarity rating for each of the 118 job characteristics. Pairs of characteristics were selected which possessed approximately the same desirability ratings and substantially different mean similarity ratings. The purpose was to produce a forced-choice rating scale in which a pair of characteristics would have equal social desirability but would differ in similarity to the job; in other words, one characteristic had a higher level of appropriateness to the job than the other. The final set contained 59 paired characteristics, as each characteristic was matched with another characteristic from the original pool of 118.

Another set of approximately 140 incumbent Border Patrol Agents filled out a Job Characteristic Similarity Survey - this was a forced-choice questionnaire containing the 59 matched characteristics. The incumbents were first asked "to choose which of two jobs (Job A or Job B) is the most similar to your own job" and second "indicate how much more similar the chosen job (A or B) is to your job than the other job is to your job." Underneath each job characteristic, there were three possible ratings: "Slightly More Similar, More Similar, and Much More Similar," thus making up a 6-point scale. The distribution of ratings for each item on this similarity survey resulted in the normative profile of the Border Patrol job that formed the scoring key for the expected-motivation questionnaire.

The resulting data on the 59 paired characteristics also indicated whether or not an item was to be considered acceptable for use in the questionnaire for applicants. If one of the matched pairs was chosen over the other matched pair by at least 60% of the incuments, the item was considered valid - otherwise eliminated. Twelve were eliminated.

The remaining 47 paired characteristics formed the basis for the 47 items that constitute the EMQ. On this questionnaire, applicants are asked to "indicate which of the jobs (Job A or Job B) you personally would prefer and how much you prefer this job over the other job: either Slightly Prefer, Prefer, or Strongly Prefer." The overall scale for each item is a 6-point scale (composed of two 3-point scales). The scale is essentially identical to the one given to incumbents (in the previous similarity survey) with the exception that the word "Prefer" is substituted for "More Similar."

The frequency distribution and the modal response(s) of the incumbent ratings on the similarity survey formed the basis for the scoring of the EMQ. The scoring procedure was based on a profile similarity algorithm that accounted for the degree of discrepancy between an applicant's response to an item and the modal response(s) given by the incumbents.

The Miniature Training and Evaluation Test is a method in which applicants are trained to perform a relevant sample of tasks for a given job and are evaluated on their ability to learn and perform these tasks upon immediate completion of training. The approach is based on the notion that applicants who can demonstrate the ability to learn and perform on the work sample can learn and perform the job with appropriate on-the-job training. This test was chosen to assess the applicants on the following two KSAOs: 1) ability to apply service policies and procedures, and 2) ability to prepare and complete accurate and concise reports in a timely manner, displaying proper written communication skills. In order to assess these abilities, it was decided that applicants would learn service policies and procedures regarding the preparation of reports and would apply these procedures in writing a report.

Subject matter experts were asked to write a simplified and condensed version of the training manual that pertained to service policies and procedures regarding the preparation of official reports. Then they were asked to develop test situations with their solutions. The solutions contained the various requirements necessary for an accurate and complete report such as: the correct form to be executed, the date and place of apprehension, the statute(s) violated, and so forth.

The Interview. Over the last several decades, research and reviews of studies regarding evidence of reliability and validity have been quite disappointing; the conclusion is that the interview has generally been ineffective in accomplishing its selection purpose. The reviews indicate the need for more structure and standardization, simplified judgment tasks, a carefully planned procedure for processing and combining information, and extensive rater training.

The following procedures were incorporated into the development of this interview. First, applicants were to be assessed on four important KSAOs. They were as follows: 1) the ability to make correct decisions quickly, under pressure, using sound judgment, 2) the ability to recognize and analyze problems and determine correct solutions, 3) the ability to recognize, evaluate, and maintain control in dangerous or highly adverse conditions, and 4) the ability to interact and work with fellow employees and deal effectively and tactfully with the public. The first three abilities would be assessed by the applicant's responses to hypothetical situations while the remaining KSAO regarding interpersonal skills would be evaluated by observing behavior during the interview. There were various reasons for retaining the hypothetical situation format. One is based on an underlying assumption of goal setting theory that intentions are related to behavior. If what people say correlates with what they do, responses to critical situations should be both informative and valid. One critical difference between the old and new interview format was that only hypothetical situations related to the three abilities would be used.

A panel of subject matter experts reviewed the original situations and retained only those that had the potential to distinguish between poor and good performance and that directly assessed only one of the three KSAOs. The second concern of the review panel was the development of benchmark responses for each hypothetical situation to assist and simplify the judgmental task of the raters. Subject matter experts independently developed both high and low quality responses to each hypothetical situation, which were reviewed by a panel of experts. Third, it was decided that more structure and standardization was necessary in the following two processes: 1) conducting the interview, and 2) determining the ratings. This was accomplished through the development of a training and interviewer guide and subsequently a training workshop.

Ratings are determined through the implementation of the following process. One panel member records the applicant's responses and behaviors. Following the interview, the notetaker reads the questions and the applicant's responses. The panel members then independently rate the applicant on a 5-point scale on each of the KSAOs. If there is disagreement among raters, a discussion follows until a final rating is agreed upon. The applicant's score is a total number of points from all the KSAOs. Finally, a cut-off score determines whether the applicant passes this stage of the selection process.

PAPER SESSION

Measuring Productivity

Chair: Susan Biesele, Salt Lake County Office of Personnel Management

Development of a Programming Productivity Measurement System

Marianne Bays, Prudential Insurance Company

The Prudential Insurance Company of America's Computer Systems and Services Office (CSSO) undertook a productivity measurement system development project in early 1982. The basic measurement problem faced was how to value programming project output or the amount of programming work accomplished. The CSSO began working with a task force of eleven project leaders who represented different applications programming areas in the organization. The group's purpose was to develop a method of assigning output values to programming projects which can then be related to the resources (# of work hours) used to generate that output in order to derive a programming productivity index.

The aim was to develop an output valuation method which:

- Was applicable to both maintenance and new development projects.
- Could be used with objectivity and consistency across CSSO.
- Could provide a valid basis for productivity analysis of applications programming areas (not individuals) over time.

A series of meetings were held which focused on identification and analysis of the factors which make one programming project more difficult or complex than another.

The system developed is called the Project Complexity Analysis System (or P-CANS). The P-CANS checklist has been tested for interrater reliability and for validity with very positive results. Several revisions have been made to the draft checklist which was tested in order to further increase the consistency with which it can be used. A four month P-CANS pilot in ten project areas of CSSO has also been completed.

Brainstorming sessions were held by eleven project leaders to develop a list of programming project complexity factors. This process generated a total of 77 factors which were grouped into rough categories. The group members independently rated each factor's applicability to new development and/or maintenance projects in the areas they represented. Project leader's ratings of applicability and variation of factors within their areas were tallied. Results were then reviewed and statistical criteria were established for calling a factor "most useful."

The group was then led in a discussion about how to best scale the factors within each cluster. The researchers took notes on this discussion which were later used as the basis for drafting a set of tentative rating scales. The next step in the project was to obtain data on each factor cluster's comparative contribution to project complexity so that cluster weights could be established. Each project leader was given a set of the draft rating scales for each factor cluster and was asked to independently rank them on the basis of the degree to which they contributed to programming project complexity in the functional area they represented. This produced eleven factor clusters which applied to new development projects ranked from 1 to 11, with 11 representing the greatest contribution to complexity. A twelfth cluster, Technical Maintenance Considerations, was included in the ranking procedure for maintenance projects.

An interrater reliability study was conducted on the draft P-CANS checklist to determine the consistency with which it could be used to value programming project output. Essentially, this was designed to be a test of how well P-CANS ratings generated by independent raters on the same projects would match. Inconsistencies in rating would help us to identify any items on the checklist which were ambiguous or poorly defined and any instructions which needed clarification. While the overall interrater reliability coefficient obtained (.983 at the .001 level of significance) was high enough to conclude that the P-CANS checklist could be used with a good degree of confidence in its measurement stability, weaknesses were noted in several of the subscales. An examination of the raw data resulted in observation of several circumstances where revisions in wording, simplification of multiple choice formats and clarifying comments could increase item reliability.

A criterion-related validity study was conducted. All correlation coefficient obtained were high enough to allow conclusion that the complexity factors and weights incorporated in the rating checklist were good approximations of the judgments that our experienced Senior Manager sample held with regard to what contributes to project complexity.

A four-month pilot of P-CANS began on December 1, 1982. Ten project areas, representing different programming sections of CSSO, participated. The purpose of the pilot was to determine how P-CANS works when applied to actual projects; to gather information which would allow us to refine the checklist items and instructions to give them better reliability and validity; and to develop a strategy for implementation of P-CANS CSSO-wide. Pilot participants' reports were encouraging. For the most part, they felt that the indices and P-CANS values were accurate reflections of comparative levels of project progress and complexity. The initial purpose of developing P-CANS was to provide area productivity indices for the monitoring of productivity from one period to another. We also plan to use the indices to demonstrate CSSO productivity trends. We also believe that on-going analyses of the indices can help us to determine if certain work habits or standards used in one area would account for higher productivity there than in other areas. We might, for example, see that an area that uses a structured life cycle methodology with defined deliverables has higher productivity than an area that plans their projects but does not specify sign off points or deliverables. We also hope to be able to use the indices to demonstrate any improvement in productivity that might be gained from using tools like application generators on projects. The pilot participants have been very positive about the use of the checklist and the information that has been derived.

Development and Use of a Computer-Assisted (SAS) Goal-Setting System: Implications for White Collar Productivity

Nicholas F. Horney, Salt River Project

The introduction of a participative work team concept (typically discussed in the organizational development literature as quality circles) to an organization can be costly due to the training required and typical use of external consultants. Furthermore, such an approach to productivity improvement may require substantial change in the organization's philosophy of worker involvement. Therefore, it would seem more appropriate for organizations to take a close look at existing organizational systems to explore what contributions they could make to productivity improvement. One such example is the performance appraisal system which also incorporated goal setting as a part of it.

One method of improving productivity in an organization is by improving upon (or establishing) a viable goal-setting process. This proposal reviews the development of the research model to be used in the study. It reflects the contributions of the goal setting and MBO literature, the management control system literature, and issues relating to white-collar productivity.

The majority of the literature on organizational applications of goal setting originates from the work of Locke (1968). Locke's principal theoretical statement, drawing from Ryan's (1970) work on intentional behavior, is that the immediate cause of behavior is task goals or behavioral intentions ("what the individual is consciously trying to do"), not the mechanical operation of reinforcers (Locke, 1968). Locke's formulation of a goal is what a person is trying to do. Therefore, a person's performance is regulated by his/her goals. Locke's theory states that hard goals provide more improvement than easy, and specifically more than general or "do-our-best" goals. The most recent review of Locke's theory and its application is incorporated in a review of goal setting from 1969 to 1980 (Locke, Shaw, Saari, and Latham, 1981). According to this review, "goal setting is most likely to improve task performance when the goals are specific and sufficiently challenging, the subjects have sufficient ability (and ability differences are controlled), feedback is provided to show progress in relation to the goal, rewards such as money are given for goal attainment, the experimenter or manager is supportive, and assigned goals are accepted by the individual."

Goal specificity appeared to be the most important aspect of the MBO process for performance improvement. However, few studies have dealt specifically with the effects of variation in goal specificity. Most of the studies cited in the reviews defined goal specificity as goals versus no goals. Reviews of the goal-setting and MBO literature appear to have one conclusion in common, that one of the major reasons goal setting procedures work is that they serve to make explicit specific performance goals.

Since MBO and goal setting procedures have been shown to increase productivity, they appear to be useful organizational development techniques. These procedures allow supervisors and subordinates to reach consensus at the cognitive level regarding the direction the subordinate should be working and spending his/her time.

Management Control Systems. Organizations contain many management control systems. These include those whose purpose is for performance appraisals and subsequent merit increases. Others exist primarily to provide ongoing feedback to employees about how they are performing their jobs.

The most important positive function a control system can perform for an individual is to give feedback about task performance. Resistance to control systems is most likely to be present when 1) the control system measures performance in a new area, 2) the control system replaces a system that people have a high investment in maintaining, 3) the standards are set without participation, and 4) the results from the control system are not fed back to the people whose performance is measured.

White Collar Productivity. A measure of productivity is any ratio of output to one or more corresponding inputs. An organization can achieve a complete productivity measurement system in one of two ways. First, it can develop a number of measures, at least one for each of the four major input factors. If all are monitored, the effects of the trade-offs can be seen by the relative changes in the different ratios.

The second approach is to implement an integrated productivity measurement system that accounts for all four partials and combines them into a total productivity measure for the organization.

Finally, illustrations of output were described which can be obtained from the computer-assisted goal-setting system. The examples demonstrated how the system could be used to plan and manage test validation and other personnel research projects.

INVITED ADDRESS

Alternative Uses of Traditional Selection Procedures

James L. Outtz, Howard University College of Nursing/Law

Among the important tasks employers must accomplish are (a) adherence to the mandates of merit selection, (b) demonstrating compliance with Title VII of the Civil Rights Act. Accomplishing the former often appears antithetical to the latter. In no area is this apparent paradox more evident than with regard to the issue of alternative selection procedures. Under the Uniform Guidelines an employer must demonstrate the validity of a selection procedure that has an adverse impact on a protected class. Moreover, the study must include an investigation of alternative selection procedures or alternative methods of using the selection procedure that would have substantially equal validity. Since traditional selection procedures rely heavily upon the written test, it is important to consider the issue of alternatives from the standpoint of alternative methods of using this selection procedure. Most, if not all, of the published research concerning alternative selection procedures, addresses the problem from the standpoint of finding suitable substitutes for paper and pencil tests.

Little, if any, research has been published concerning the possibility of identifying alternative methods of using a selection procedure that would reduce adverse impact without reducing validity. This is particularly puzzling in light of recent professional standards which define validity as the accuracy of inferences drawn from scores on a test. Thus, validity has less to do with the properties of a test than with the manner in which the test is used. In light of the professional and legal obligation to pursue alternative methods of using selection procedures, I would like to discuss several alternative methods which appear promising. The traditional procedure in the law enforcement area which I shall use in my discussions is based upon a multiple hurdles model wherein each applicant or candidate for promotion must successfully negotiate a series of selection or promotion devices. The first hurdle is usually a paper-and-pencil test. Subsequent selection devices may include a background investigation, physical agility test, psychological evaluation, physical examination or oral interview, etc. The key component is the written test and the manner in which it is used.

Test Content. Test content constitutes a factor of consideration with regard to alternative methods of using a test. One study shows the degree of adverse impact for the various components of a test for the position of State Trooper (eastern United States). The test was administered in a concurrent validation study to a sample of 33 minority and 87 nonminority persons who had entered the states' training academy. Fifteen subtests were arranged into three batteries of five tests each based on the degree of adverse impact of the tests within each battery. Validity coefficients for each battery were calculated based upon four different criterion measures: supervisor's rating, average productivity/activity, final academy average, and average number of arrests.

Results suggest (a) that the best predictor group depends upon the criterion predicted and (b) the low adverse impact battery produced validity coefficients comparable to those for the high adverse impact battery for every criterion except academy performance.

In another jurisdiction in the eastern United States the employer revised an entry-level test for state trooper by deleting items from the originally constructed test based primarily upon three decision rules: (a) high adverse impact, (b) low item validity and (c) high item difficulty. In a concurrent validation study (N=246) the original test produced a validity coefficient of .27 with a criterion of supervisor ratings. The validation sample of 246 was then divided into a follow-up sample of 123 and a cross-validation sample of 121. Rescoring of the test based upon the shorter number of items after the deletions produced a validity coefficient of .40 in the followup sample and .37 in the cross-validation sample. The revised test was subsequently administered to 2,213 non-minority and 244 minority applicants seeking employment as a State Trooper. Further item analyses were done on the applicant group, cataloguing each test item on the basis of a number of criteria including degree of adverse impact, item total score correlation, and correlation with self-ratings made by the applicants at the time they took the test. Based upon these criteria, several shortened versions of the test were created and the validity of each assessed by rescoring the answer sheets of the concurrent validation sample. A final version of the test chosen for use significantly improved the pass rate for minority applicants.

Method of Test Administration. A second area which appears to have some relevance with regard to alternative methods of using a test is the manner in which the test is administered. Candidates can be given pre-test materials to study or the employer may elect to conduct pre-test orientation sessions for all applicants. I have found that providing pre-test information can be particularly beneficial in promotional examinations. For a promotional written test, pre-test materials typically consist of a so-called reading list considered fair game as a source for test questions. The strategies used by the candidate in utilizing the materials are very important. The cognitive style with which the candidate approaches study materials in the form of reading lists - in a field dependent or global manner or in an analytical of field independent way - may have significant effects upon the usefulness of the study materials. I would propose that cognitive style may well be one of the most potent yet least recognized moderators in selection research.

Although public employers may not be able to incorporate such time consuming or complex instructional systems into the nuts and bolts milieu of employment selection, there are innovations to be considered. It may be possible to reduce the guessing game mentality that surrounds candidate preparation for promotional examinations. Samples can be provided of 1800 open ended questions that were formulated to assist police officers in identifying critical concepts within source material assigned for study in connection with a police sergeant promotional examination. Thus, the open ended items serve as a study guide. Other alternatives to traditional examination administration procedures include reducing the overall amount of reading material assigned by testing only on those areas most critical to job success or allowing candidates access to source material while taking the test (open book procedure) if they would normally have to refer to such information in order to successfully perform the job.

In my experience, non-traditional examination administration procedures offer promise in the reduction of nonminority - minority differences without loss of validities.

Interpretation of Examination Scores. In the traditional selection procedure applicants are rank ordered on the basis of a written test score, often with the most trivial differences in scores considered meaningful. As early as 1978, courts were finding fault with the use of a written test to rank order applicants on the basis of small test score differences. It would appear to me that an appropriate alternative would be to develop test score bands based upon the standard error of measurement. Applicants whose test scores fall within a given band would be considered equally qualified; however, applicants in higher bands would be considered more qualified than those in lower bands. Since all applicants within a band would by definition be equally qualified based upon the written test, all would move on to the next hurdle in the selection process. Under this system, the number of minority applicants qualifying for the next hurdle would be larger than otherwise would be the case, yet the validity of the test would not be compromised.

Summary. I submit that the primary challenge of the 80's with regard to equal employment opportunity should not be only to find alternative selection procedures but also to identify alternative methods by which valid procedures can be used.

WINNER: STUDENT PAPER COMPETITION

Chair: Marianne Bays, The Prudential Insurance Company of America

Student Award Winner: Dennis Doverspike, University of Nebraska at Omaha

A Statistical Analysis of Internal Sex Bias In A Job Evaluation Instrument

Dennis Doverspike, University of Nebraska at Omaha

This study was a preliminary investigation into the utility of internal bias analysis for the study of sex bias in job evaluation instruments. It involved the adaptation of techniques from the study of bias in aptitude tests to a current, applied problem in personnel management.

Research Problem. As a result of comparable worth theory, job evaluation has emerged as a topic of considerable interest in the psychological and legal literatures. Comparable worth theory required a measure of job worth. Traditionally, job evaluation instruments have provided a measure of job worth.

Both proponents and opponents of comparable worth alleged, however, that job evaluation might be inappropriate for the measurement of worth. One proponent's argument held that job evaluation instruments, not the raters but the internal structure of the instrument, were biased against female jobs. That is, job evaluation instruments differentiated or measured worth more adequately for male jobs.

Of course, proponents of comparable worth believed that pay was sex biased. Therefore, no external criterion existed for a test of sex bias. In the aptitude test literature, internal test bias analysis methods have been developed for the situation in which no external criterion existed.

It was proposed that the question of sex bias was one of similarity of psychometric properties of job evaluation instruments across sex of jobs. Therefore, the methods of internal bias analysis were applicable to the analysis of sex bias in job evaluation instruments.

Review of Literature. A comprehensive review of the job evaluation literature indicated that no study was found which investigated sex bias in job evaluation using the internal bias method.

The literature review revealed that sex bias was unlikely. The findings of studies were consistent across varying job types. Job evaluation ratings were also reliable and valid. The use of ratings of job descriptions was supported in the literature and was a common methodology. Comparable worth theory and internal bias analysis were also reviewed. There were a number of ambiguities in comparable worth theory that made scientific study difficult.

Internal bias methods were based on construct validity. Internal bias analysis techniques include investigations of differences by group in reliability, factor structure, scale - total score correlations, and mean differences of scales.

Methodology. Four upper level, trained graduate students rated 105 male and 105 female job descriptions. The jobs were identified using census data and an 80% criterion. Jobs were randomly sampled from office occupations. Job descriptions were adapted from the Dictionary of Occupational Titles (DOT). In addition to the job evaluation ratings, job data were obtained from the DOT.

Ratings were made on a 15 factor point method job evaluation instrument, the Comprehensive Job Evaluation Technique (CJET). The instrument consisted of ten traditional and five nontraditional factors and was developed based on the literature review.

Data Analysis. The results were analyzed based on the internal bias analysis method. Statistical analyses included differences in generalizability coefficients by sex, a scale by sex by job ANOVA, factor analysis by sex, differences in scale - total correlations, and partial correlations between sex of job and scales controlling for total point score.

Findings. Inspection of the statistical analysis of the CJET revealed that ratings were reliable and valid. Overall, the CJET differentiated female jobs as well as or better than male jobs. Scales which were biased in favor of male jobs included education, financial responsibility, surroundings and hazards. Scales which were biased in favor of female jobs included previous experience, visual attention, physical effort, supervisory responsibility, responsibility for the safety of others, consulting, and monotony.

Based on the results, the future development of an unbiased instrument depends on a proper balancing of scales measuring interactions with people and scales measuring interactions with things. Better measures are needed of interactions with things, particularly data manipulation. Three of the nontraditional scales added little to the CJET and were negatively related to worth within female jobs. This illustrates the problem with simple subjective assessments of job evaluation scales. There was some agreement among the internal bias methods and their application to job evaluation was accomplished without any major problems. Internal bias methods involve the comparison of bias of prediction in construct validity. The major problem is a tradeoff between bias and fairness.

Limitations and extensions of internal bias analysis were discussed. The major untested assumption involved the issue of bias as mean differences versus bias as errors of prediction. Another problem was the sampling of jobs, which excluded factory jobs. Extensions included the use of internal bias analysis to investigate bias in selection instruments and performance appraisal. Further research was suggested on the relationship of perceptions of pay to comparable worth and job evaluation.

Conclusion. Individual job evaluation scales were biased in favor of both male and female jobs. However, both the present study and the review of the literature indicated that, overall, the validity and reliability of job evaluation instruments as measures of worth are not dependent on the sex of the jobs being rated. Internal bias methods are applicable to the future study of bias in job evaluation instruments.

PRESENTATIONS BY STUDENT PAPER COMPETITION SEMI-FINALISTS

An Assessment of a Measure of Job Satisfaction in the Philippines and the United States

Maria Veronica de Vera, University of Illinois, Urbana-Champaign

Overview. Measuring instruments developed in western environments have frequently been imported for use in other cultural settings. One should be cautious, however, in accepting the validity of a test without question. A test that looks good does not mean it is an unbiased measure.

This investigation focused on the relevance of the construct of job satisfaction in the Philippine and American cultures. Determining whether job satisfaction is perceived in the same manner in the Philippines as in the United States can only be checked if one administers the same measuring instrument to both cultures. The large number of bilingual workers in the Philippines creates a unique opportunity for the study of independent effects of translation and culture on measured job satisfaction.

The Job Descriptive Index (JDI) developed by Smith, Kendall, and Hulin (1969) was used in this inquiry. Achieving a "culture-fair" translation of the JDI will make theoretical and applied contributions to the field of international personnel assessment. It will provide the Philippines and other international organizations who employ Filipinos a psychometrically sound measure of job satisfaction. It will also allow for meaningful comparisons which may contribute in the determination of the universality of the construct job satisfaction.

Method and Procedures. (1) Instruments used: The original English version and a Filipino translation of the JDI was used for this investigation. (2) Translation procedure: A combination of the following techniques was used: back translation, committee approach, and pre-testing. The JDI was translated into Filipino by four bilinguals working independently. A different set of bilinguals back-translated the Filipino version of the JDI to English. The researcher, who is also a bilingual, assessed the quality of the final version of the JDI in consultation with a Filipino language professor. (3) Description of Data: Complete data were obtained from 1100 American employees, primarily from the nonmanagerial staff of a large U.S. retail organization (courtesy of Dr. C. Hulin). Complete data was also obtained from 1342 bilingual employees from various Philippine companies. Employees from all job levels were represented in this data set. The survey was administered to groups of 15 to 40. Individuals were randomly assigned to two separate groups. Group 1 completed the English version of the JDI and Group 2 completed the Filipino translation of the JDI. (4) Design and Analysis: Determining the quality of the translation of the JDI items involved several comparison of the three JDI data sets:

A - Filipino JDI items compared with English JDI items from the Filipino sample, to test the effect of translation;
B - English JDI items from the American sample with English JDI items from the Filipino sample, to test the effect of culture; and
C - Filipino JDI items with English JDI items from the American sample, to test the cumulative effect of translation and culture.

This investigation utilized Item Response Theory (IRT) approaches in analyzing the data.

Findings and Conclusions. With the aid of the IRT approaches to statistical analysis, biased and unbiased translated JDI items were detected. Results from Lord's chi-square test indicated that only 30% of the translated JDI items were found to be unbiased. This conclusion was based on the similarity of the Item Characteristic Curve (ICCs). The F-test yielded similar results.

There were instances, however, where ICCs of a particular item were found to be significantly different for both comparisons B and C but not significantly different in comparison A. On these items it appears that, although appropriate translations were achieved, the translated item conveys a different empirical meaning from that of the American item. This finding demonstrated the need to consider cultural, as well as language, issues in test translation.

It is apparent that establishing equivalence of scales in two different languages and cultures is difficult. The present study shows that: 1) the use of measurements in different cultures without correcting for biased items may result in misleading conclusions; and 2) adequate literal translation of words does not guarantee equivalence in empirical meaning.

Subsequent research should focus on validating the revised JDI scale in the Philippine culture and on developing other measures of job satisfaction specifically for the Philippine culture.

Individuality in the Organizational Context: Conceptual Approach and Its Applications

Michael D. Mumford, University of Georgia

This paper represents an attempt to formulate an efficient and general model for the definition of individuality as well as to elaborate some of its potential applications. It is suggested that homogeneous subgroups of similar individuals can be identified on the basis of their life history. These subgroups may be viewed as a sort of model individual and known characteristics of subgroup members can form a basis for the study and control of individuality.

Previous research at the University of Georgia has been based on a single assumption; that is, while the behavioral variation among individuals is enormous, it is not infinite and so some individuals should be more similar to one another than they are to other individuals across time in a wide variety of behaviors and experiences. In these classification studies, we have attempted to obtain a comprehensive description of the individual's behaviors and experiences, and then a clustering algorithm has been employed in order to identify natural occurring subgroups, or clusters, of individuals in the population at large. Biodata or background data items were chosen as a primary vehicle for this research effort. A 118-item biodata form, termed the Biodata Questionnaire (BQ), was constructed and designed to cover significant behaviors and experiences occurring during the adolescent and childhood period. It was administered during freshman orientation to a sample of some 10,000 male and female undergraduates entering the University of Georgia between 1968 and 1974. It is of note, that each entering class or cohort provided roughly 2,000 subjects. As the members of each cohort graduated from the University, they were mailed a second biodata questionnaire termed the College Experience Inventory (CEI) focusing on significant behaviors and experiences during the college years.

In the data analysis, the male and female item responses obtained from each instrument were intercorrelated and summarized through an orthogonal principal of components analysis. The principal components dimensions obtained for each sex group on this instrument were then labeled on the basis of the items yielding loadings above .30, and scores on these components were obtained for all of the subjects.

Within each sex group, the scores obtained for the subjects on a given instrument's dimensions were used to define similarity by forming a set of profiles. The similarity between a certain subject's profile and profiles of the other members of his/her sex group was then assessed through Cronbach d^2 measure of profile similarity. The d^2 index is sensitive to similarity in the shape, elevation, and scatter profiles, and the resulting matrix of similarity data subsequently formed the foundation for a set of Ward and Hook clusterings.

In subgrouping the data obtained from each instrument, roughly 10 to 15 subgroups were identified within the male and female samples. The subgroups identified in the analysis of each instrument were associated with a very broad, but interpretable pattern of behaviors and experiences which differentiated subgroup members from sample as a whole on the relevent items and summary dimensions. Also, the subgroups or classes appeared to be stable across changes in sample structure in the sense that they could be applied in new samples with little loss in the accuracy of classification. Subgroup status proved to be a highly effective predictor of real world criteria; in fact, it typically yielded better prediction than the more traditional trait measures.

A detailed examination of these protocols indicated that each subgroup was associated with a unique and interpretable pattern of behaviors and experiences over time. In sum, the composite subgroups appeared to provide an adequate description and classification of individuality over time.

Further inspection of the subgroup differential characteristic yielded a number of important results bearing on the nature of the individuality and the subgroups. First, subgroups which displayed very different patterns of behaviors and experiences over time might, at a single point in time, display nearly identical patterns. The empirical existence of this local convergence phenomena indicates that individuality can only be adequately defined on a cross-time basis. Second, each subgroup was associated with, non-zero, statistically significant differences in the relationships among the summary dimensions which were readily interpretable, given knowledge of the subgroups' other differential characteristics. This result indicated that the subgroups were characterized by qualitative or organizational differences, as well as quantitative differences. Third, certain differential characteristics appeared to exert a preponderant control over the later behaviors and experiences of subgroup members. For instance, one subgroup of high-ability females did not do as well on the job, as might be expected, due to a focus on nurturant activities. Fourth, the later implications or meaning of a given behavior or experience differed across the subgroups as a function of a variety of complex interactions. An example of this phenomena may be found in collegiate rebellion which was associated with favorable outcomes in adulthood when individuals were raised in a poor home environment, but unfavorable outcomes when individuals were raised in a good home environment. Fifth, the differential characteristics of each subgroup appeared to be centered around a particular mode of adaption favored by subgroup members. Finally, the characteristic behaviors and experiences of subgroup members changed in a systematic fashion over time, as a result of new adaptive demands, in new environments.

INVITED ADDRESS

Recent Developments in Employment Law

Lawrence Z. Lorber, Attorney; Breed, Abbott and Morgan, Washington, D.C.

In reviewing EEO cases of the last eight months in preparing for this talk, and considering new developments and focusing on the areas the courts have been addressing, I come to you with the observation that the new developments of the law of the last year have not been very new and have not been very revolutionary. I offer that observation at a time when we keep on reading in the newspapers that we are witnessing a sea of change in the view towards equal employment law and towards the regulation of the work place of which equal employment law is probably the main area.

The first point that msut be emphasized and re-emphasized is that the legal regulation of the workplace -- the regulations and the law that you all have to deal with and translate from arcane jurisprudential writing into practical reality--is a world that is marked almost uniquely by a split and a diffusion of authority and responsiblity. As you all know, as employers you could have the unique pleasure of being sued in federal or state court by the government, federal, state, or local, or by private litigants, under a host of requirements, laws, and statutes. The point is that there is no one single focus for the development of equal employment law. And to the extent that there is a focus in my view it has been set almost uniquely by private litigants raising issues and raising concerns under statutory law over the past ten or twelve years rather than by the Government. And indeed in viewing some of the major cases that we are facing today -- some of the major equal employment cases which have reached the headlines, the New Orleans police case, the Detroit police case, the Boston police case -- we find that those cases have been brought and developed initially by private litigants and that the federal government and the federal government's own view of what the law requires is a view that is not necessarily adhered to by the courts nor is it necessarily followed by private litigants or persons who have an interest in equal employment. Which is simply to say that regardless of the rhetoric one hears from Washington, the reality of equal employment law, in my view, is that it has not changed substantively over the last year or year-and-a-half. The remedies available to litigants who have brought an action and proved a case of discrimination remain the same. The courts, certainly since 1971 and indeed since before that time, have not been loath where the facts warranted to order remedies which might indeed impinge upon or impair the expectations of persons as to what their job futures or possibilities are at the expense and for the benefits of other persons who, though they may not themselves have suffered discrimination, are part of a group or class which the facts, the law, the statistics have shown to have suffered exclusion from the workplace. In other words, what is bandied about as preferences, quotas, even reverse discrimination, does still exist in the law. The rhetoric we hear out of Washington saying that there will be no more preferences, no more goals, no more activities which impinge upon and offer preferences to persons who have not individually suffered discrimination, in my view at this time, is simply rhetoric and I believe will remain rhetoric at least as long as the judiciary is constituted as it currently is.

(Mr. Lorber's complete address presented reviews of recent cases, court findings and decisions, and federal government agency actions related to these statements and his ensuing remarks, which are not reproduced in these Proceedings.)

Now I shall get into areas where I think personnel professionals need to focus. What has been interesting this year in the development of the law is a move towards a refinement of the law of equal employment and, for the first time, a direction and an examination of the proper application of commonly accepted equal employment principles in the area of higher professional jobs, white collar jobs, and jobs where there is clearly an accepted need for requisite skills for persons to perform that job. Now it is certainly not to say that the courts are fashioning a new view of the law. They are not. But what is happening in the type of cases in what I will call the white collar area, the professional area, is that these cases which often involve claims of age discrimination, compensation discrimination, promotion or hiring discrimination or in the common law area related to all of these, the courts have seemed to be accepting law already established, viewing the higher level jobs as worthy and necessary to be judged on a higher standard. The Pounce v. Prudential case in the Fifth Circuit, a private sector case, is one example where the court, having gone through a typical statistical analysis, did point out and refuse to accept the blanket assertion that the general statistical showings of general availability for minorities, without reference to the skill levels necessary to perform the jobs, would be sufficient in higher level jobs. The court, in that case, would not blindly and blankly accept those statistics as sufficient to prove a case of discrimination. Rather, the court looked to the specific job requirements, the accepted job requirements and the shown job requirements, and said that the plaintiffs had a higher burden they had to rebut, and they had to answer not only the statistics but they had to show that either their class of plaintiffs met the criteria necessary for the job or that the criteria itself was designed in an exclusionary manner. And in so fashioning that type of decision, the court put a burden at least upon plaintiffs to meet the higher burden of the disparate treatment case and to show that the employer intended in individual instances to discriminate. It did not accept as a premise that a policy, which perhaps resulted in a statistical exclusion of minorities, was indeed a neutral across-the-board policy to which an adverse impact analysis might be implied. The Supreme Court said in Burdine that the plaintiff had to meet a higher burden and show that the plaintiff should have been selected and was, indeed, more qualified than the person that was selected. In a lot of instances the courts are starting to erode the distinctions. The Fifth Circuit in the Uncle Ben's case did state and did hold that Burdine and that type of analysis would not apply to a classic adverse impact case; that in a classical impact case, perhaps a testing or promotion case, that you still had the Griggs type of analysis and that the burden on the employer was significant. However, Uncle Ben did not involve higher level employees and I think the trend that we are seeing is that the type of burden being put on the plaintiff is a greater one for higher level employees and that the plaintiff has to show more than a statistical problem. On the other hand, what is also happening to equalize what I view as a subtle but nevertheless significant change in the area of compensation, in the area of age discrimination, in the area of termination at will, in the classical area of Title VII, sex and race promotion or demotion/termination cases, is that employers have to offer in both adverse impact and disparate treatment cases, professionally acceptable job analysis, quantification of the functions that the job requires, a quantification and a defensible analysis, and a review of the employee's work performance. The courts are now focusing on refining the law. The easy cases are now fading away and the massive cases where there is an industrywide or employerwide exclusion showing statistical differences between the participation rate of minorities or women is out of line with what one would expect and with what the statistics might show to be their availability. Those types of cases, as I put it--the easy cases, the big cases, the granddaddy of all type of class action cases--are simply withering away, if for no other reason than litigation already touching a lot of employers. Currently, the focus on better jobs combined with a new class of plaintiff who individually has the knowledge of the employer's own employment practices, the knowledge of their legal rights, the ability to pursue a litigation without relying on a larger class, and an understanding that the law is now opening up a host of avenues where individualized grievances over employment treatment issues--brought in the union context under some sort of a collective bargaining, or individual grievance system--are now having recourse to the courts to allege that their employment/career opportunities have been hindered. And what is happening, is that a focus on the employer's individual personnel systems, the job analysis for individual jobs, the type of functions and requirements which the employer expects from an individual who is performing any one of a number of jobs, and at the same time, the employer's own actions, the perogatives of the employer, the ability to act in a manner which the employer believes is right, is being challenged. The employer that never took the time nor the effort to document actions which might result in adverse termination or employment changes is losing in court. That type of focus is now becoming more and more prevalent so that the legal mandates and the legal imperatives on the personnel profession are directing itself not to be broad based to make sure the numbers are right--but are now focusing on what is perhaps the area where it should have always focused, on making sure the subjective, arbitrary, abusive type of personnel functions are not carried out. I would offer an observation that it does not seem to be a proper function of law to tell an employer to be a smart employer, to be a good employer. A quote I always use is by Bob Guion: "There is nothing that says an employer can't be fairly stupid as long as that employer is stupid fairly." That theory is being challenged, and so too, personnel is being challenged. The legal challenge of the 80's will be to develop a fair, sound and accurate personnel system. Perhaps overly bureaucratized and hamstrung by the fear of a lawsuit, but probably better in the long run.

INVITED SPEAKER

The Impact of Selection on Workforce Productivity and Output

Frank L. Schmidt, U.S. Office of Personnel Management and
George Washington University

This presentation addressed two basic ways to improve productivity: (1) through technical improvements (e.g., robots, word processors, etc.) and (2) by improving the job performance of employees. The first represents a focus on technology, the second, a focus on people. The second route divides into two basic approaches: (1) select better workers to begin with and (2) improve the management of present employees (e.g., through training, work motivation programs, etc.). Methods for determining the economic value of technological improvements have only recently incorporated the economic impact of selection and nonselection organizational interventions of a psychological nature (See April 1983 American Psychologist).

This presentation discussed methods for determining the economic value of selection interventions. To evaluate the dollar value of improved job performance levels resulting from selection, one must be able to estimate the standard deviation of job performance in dollar terms (SD_y). This s the difference in dollars between the yearly output of goods and services of the average employee and an employee at the 85th percentile of job performance. A method of estimating this parameter based on expert judgment by experienced supervisors was described. Applications of this method indicate SD_y is quite large-from 40% to 60% of annual salary and often in the $8,000 to $11,000 range. Examples were presented for a typical job of increased output in dollars of new hires resulting from improved selection for various combinations of selection ratio (SR) and validity increase (Δr). These gains are substantial.

The results of a meta-analysis of 29 studies were reported which showed that the standard deviation of employee output as a percentage of mean output averages 20%, with little variance around this mean. This indicates an employee at the 85th percentile typically produces 20% more than the average employee. This finding allows us to compute gains from selection as percentage increases in output. Examples of such percentage increases were presented for a typical job for various combinations of SR and Δr. Percentage increases in output that might be considered small by many psychologists (e.g., 5% - 8%) were shown to represent large dollar values.

Organizations with a fixed amount of total output that they want to produce can take the gain from improved selection in the form of reduced hiring and smaller workforces, resulting in reduced payroll costs. Examples of this approach to utility calculation were given. These gains were also substantial.

Failure to hire from the top down based on test scores results in greatly reduced selection utility. For example, use of the minimum cutoff method often results in a loss of 80% - 90% of selection utility. If the goal is reduced adverse impact, it is far better to hire from the top down separately within each ethnic group. This procedure usually results in a 10% - 15% loss in utility. Reduced selection utility may be one cause of the decline in productivity growth in the U.S.

PAPER SESSION

Physical Ability Testing

Chair: Susan Cheatham, Federal Express Corporation

Developing and Validation of A Physical Performance Test For Screening Vermont State Police Applicants

Jay W. Wisner, Chief of Personnel Research & Examinations, State of Vermont

This project began as a result of a job analysis study of the entry-level State Police rank (Trooper Second Class) conducted in the summer of 1980. The physical agility test in use at the time this project was initiated consisted of separate sets of exercises for male and female applicants. No effort had been made to select exercises for inclusion in the test which were related to tasks performed by State Police officers, or to set standards for passing the exercises which were related to job performance requirements.

Job Analysis Method. A group of subject matter experts collected additional information necessary for the job analysis. The group reviewed the task statements which were included in the original job analysis. For each of the tasks which has a physical component, the group listed examples of the physical activity required for successful task performance. The author then developed a list of twenty-one physical tasks which summarized the examples of physical activities described by the subject matter experts. The group then proceeded to discuss methods for assessing the performance of applicants in each of the tasks. The author summarized the group's suggestions in the form of a draft set of physical performance exercises.

The next step in the project was the construction of test facilities so that a pretest of the proposed test could be conducted. A representative sample of current Troopers pretested the physical performance test. Three were female. The group was limited to individuals under 35 years old, because that is the current maximum age for applicants. In order to provide a more stable basis for setting these norms, the author requested that data on an initial group of applicants be collected, to be combined with the pretest results.

This project was designed to develop physical performance exercises which would be job related, and to avoid using calisthenic type exercises which might or might not measure underlying constructs of physical ability. However, because standard exercises have been used so frequently as indicators of physical performance, the author decided to examine the relationship between such measures and performance on the new test.

The results show statistically significant ($p < .05$) correlations between the time to complete the pursuit simulation obstacle course and the measures of sit-ups, push-ups, 50 yard dash, body fat, pull-ups, mile and one half run, bench press weight, stress test, and rope climb. All of these correlations were in the direction which would be expected. The time to complete the rescue simulation exercise showed statistically significant correlations with only three of the other measures: resting pulse rate, bench press weight, and the rope climb. The correlation with resting pulse rate was not in the expected direction; faster times for the simulation were associated with higher resting pulse rates. The correlation between the bench press weight and rescue simulation time was not surprising, since the rescue simulation involves dragging a 150 pound dummy as the major element of the exercise. The relationship to the rope climb is less easy to explain; it is possible that the two measures both depend on a combination of speed and upper body strength.

Results of the First Year Using the Original Test. A total of 132 applicants were tested in 1981. The test results indicated that the test might have had an adverse impact on female applicants. Of 125 male applicants tested, 115 passed the test, a passing rate of 92%. Seven female applicants were tested and only three passed, a passing rate of 43%. A closer look at the results showed that no applicants failed on the basis of time required to complete the exercises. All of the failures were due to two of the obstacles in the pursuit simulation exercise. Six male and three female applicants were unable to get over the six foot wall. Four male applicants and one female applicant did not achieve the seven and one-half foot "ditch jump" required.

Modifications of the Test. As a result of this experience, certain modifications of the obstacle course were made. A total of 131 applicants reached the physical performance test phase of the screening process, consisting of 120 male applicants and eleven female applicants. One-hundred and fourteen of the male applicants completed the modified test successfully, a passing rate of 95%. Nine of the female candidates also passed, for a rate of 82%. The modified test did not show adverse impact on female applicants.

Police Physical Ability Tests: Can They Ever Be Valid?

Patrick T. Maher, Personnel and Organization Development Consultants, Inc.

This expository paper discusses the many issues that remain to be resolved in relation to the use of Physical Ability Tests (PATs) in a law enforcement agency.

The premise of this paper is that the nature of the police job does not lend itself to a job-related physical ability test. Historically, prior to the implementation of the Civil Rights Act of 1964 and its application to local government in 1972, police departments generally maintained minimum height and weight standards. The theory behind these standards was that law enforcement officers need physical strength to perform their law enforcement duties, and height and weight is an indication of that strength. Once Title VII was made applicable to law enforcement, individuals began to challenge such standards as having disparate impact and being discriminatory. To replace the loss as a result of legal decisions of minimum height and weight standards, law enforcement agencies turned to physical ability (agility) tests.

In the absence of adverse impact, there is generally no legal reason to be concerned with the validation of PATs. PATs, however, carry definite adverse impact against the same protected group that was most severely disadvantaged by the height and weight standards--women. In addition to adverse impact against women, PATs appear to have adverse impact against older persons, and could thus trigger Age Discrimination in Employment Act (ADEA) issues. Currently, many police agencies, especially many that use PATs, have maximum age limits in the mid 30's. As age limits are eliminated, and recent litigation seems to indicate it is only a matter of time until they are, we can expect adverse impact based on age, and subsequent ADEA litigation.

Maher's paper reviews frequent components of physical ability tests used in recent and current public safety selection tests. His reviews are addressed to supporting his stated premise, and are expository - not reviews of specific studies nor of specific published research.

His paper emphasized the importance of recognizing the problem, inherent in requiring physical standards, standards might be acquired in probationary recruit training.

Any selection procedure can only measure entry level KSAs, not full performance KSAs. Thus, if the PAT is used as a pre-selection procedure, then the recruit training course cannot expect to train in physical fitness or in the PAT since passage of the PAT as a condition of acceptance to the police academy is sufficient proof of physical fitness to perform the critical duties of a police officer; and having once passed it, to qualify for hiring, there is no need to make additional passage of it a condition for actual employment.

PATs have significant adverse impact against women, and probably against older persons, triggering the requirements that they be shown, by professionally acceptable methods, to be predictive of or significantly correlated with important elements of work behavior which comprise or are relevant to the job of police officer.

PATs are intended to measure the physical condition of police officers and their ability to perform various motor functions, having replaced the use of height and weight requirements which at one time were believed to provide evidence that such physical ability was present.

In examining the duties of a police officer, there appears to be considerable reason to doubt that PATs are measuring critical elements, but instead are measuring peripheral elements of work behavior.

Evidence suggests that physical tasks are an in frequent occurrence on the part of police officers. They occur less than 15 percent of the time. Thus, frequency of occurrence is not a basis for determining that physical tasks are critical to the duty of police officers.

Another problem is that although PATs are based on content validation, they are not a representative sample of the content of the job. Police officers infrequently perform physical tasks alone, especially those involving the apprehension of suspects. Thus, to require a participant to perform a PAT without any assistance is inconsistent with the actual working conditions.

A pure content validation study is not a proper method to use in determining the criticality of physical tasks to the job of police officer. At a minimum, some criterion validation strategies will have to be included, and much more extensive data collection and analysis is necessary in order to substantiate that physical abilities are important elements of work behaviors for police officers in a given police department.

Conclusion. The available evidence indicates that PAT components may not be measuring critical or essential elements of work behaviors, but rather peripheral elements that cannot be defended when adverse impact exists. It may well be that PATs are no more valid than the minimum height and weight standards once so passionately defended by law enforcement and compelled to be discarded once it was determined that there was no empirical basis for defending them.

PAPER SESSION

The Development and Validation of Selection Standards for Law Enforcement

Chair: Phil Carlin, City of Tucson
Discussant: Ricki Buckly, Psychological Services, Inc.

Validity Generalization Results for Law Enforcement Occupations

Hannah Hirsh and Lois C. Northrop, U.S. Office of Personnel Management

Theoretical Background. This paper reports the application of the Schmidt-Hunter validity generalization procedure (1977) to law enforcement occupations. The procedure as designed by Schmidt & Hunter has two related purposes: (1) to test empirically the hypothesis that test validity is situationally specific, and (2) to test the degree to which validity is generalizable across different jobs and settings.

The Current Study. The current study was designed to document the validity of written tests of cognitive ability used to select law enforcement personnel, and to provide results that could guide the choice of tests to be used for new law enforcement or protective service occupations.

The occupations we focused on are those included in category 37, protective service occupations, of the Dictionary of Occupational Titles (U.S. Department of Labor, 1977), excluding firefighters and armed forces enlisted personnel. Since over 80% of the studies focused on police officers and detectives in public service, we analyzed these data separately, and then compared them to the pooled data from all covered occupations as a test of how far the validities would generalize. If the addition of coefficients from the other occupations would not significantly change the percentage of variance accounted for by artifacts, or the degree of test validity, this would serve as an empirical demonstration that narrower job types, within the broader category of law enforcement occupations, do not moderate validities.

We collected criterion related validity studies of a law enforcement occupation, testing cognitive abilities and using either a proficiency or training criterion where sample size, test type, criterion type and validity coefficient could be coded and included in our analysis. In all, we were able to use 40 studies representing a somewhat greater number of independent samples which contributed a total of 381 validity coefficients. Of these 381, 242 used a proficiency criterion and 138 used a training criterion. Since we are still analyzing our proficiency data, the results presented in this report are for training criteria only. It is gratifying to note that the average sample size of these training studies was 142, and that only 22% were at or below an N of 68.

Decision Rules. The decision rules used in treating the data in general follow those established by Schmidt, Hunter and Pearlman in their validity generalization research. I would, however, like to note that we limited our analyses to those validity distributions which contained at least 6 coefficients. We used the Schmidt-Hunter interactive procedures for all analyses.

Results. The police and detective in public service grouping contained sufficient coefficients to analyze data for verbal, quantitative, reasoning, spatial/mechanical and memory test types as well as for two composite test types. Results showed that most of the between-study variation in validities could be attributed to four statistical artifacts, and that the hypothesis of situational specificity could be definitively rejected for three of the seven distributions. In the remaining four cases, validity generalization was justifiable on the basis of the substantial values of the validities at the foot of the distribution of estimated true validities. Results of a file drawer analysis indicated that the chances that the current findings could be nullified by existing undiscovered studies were small. Detailed statistical data were presented.

Future Analyses. The next step in our data analysis will be to analyze the results for proficiency criteria for both the police and detective job family and pooled job families. We will also compute correlations between average test type validities for both types of criteria to examine the similarity of validity patterns. Additionally, we will re-analyze all the data using our empirically based artifact distributions. Our overall results will then be compared to the previous body of validity generalization literature. Our completed analyses will also be used to guide the choice of cognitive abilities tests for new law enforcement or protective service series.

A Case of Increased Selection Test Validity By Assignment Of Job Relevant Language Points

Frances Brogan, U.S. Office of Personnel Management

A two-part predictive criterion-related validity study was conducted of a cognitive abilities test for selection for a Federal law enforcement occupation. Validity coefficients were first obtained for selection test scores and scores for a training course required of all new agents. A second analysis was done using the same selection test scores, but with five points added to the selection scores of those individuals who claimed to have prior Spanish language ability. Spanish points were presumed to be relevant, because of the crucial importance of Spanish language ability for this particular occupation and the incorporation of Spanish language training in the law enforcement training program.

Individuals selected during 1978 who completed the training course served as the research participants for this validation study. Of the 191 participants, 180 were male, 11 female, and 57 classified themselves as Hispanic. Seventy individuals classified themselves as having Spanish language ability.

Predictor. A 65-item written cognitive abilities test was used in selecting applicants at the entry level for this law enforcement occupation. Competence in two areas--verbal skills and judgment--measured. These abilities were evaluated by five types of items including Reading Comprehension, Vocabulary, and English Usage items--as test of verbal abilities--and General Comprehension and Logical Order of Events items--as indicators of judgment.

Criteria. As previously mentioned, the criteria in the present study were scores from a training course required of all individuals entering this occupation. The training scores were considered to be appropriate criteria, since the training program represents an essential part of the job. The overall training score for the study sample is a composite of several area scores including Spanish (30%), Law (50%), Operational Subjects (5%), Firearms (5%), and Police Training Division Courses (10%). These scores were based largely on objective examinations. The Spanish language score was based upon examination grades in the required Spanish Training Course. An overall training score, with Spanish language score removed constituted a third criterion variable. The remaining area scores, including Law, Operational Subjects, Firearms, and Police Training Division courses, then made up this criterion score.

Correlation of the Test with Training Criteria. The relationship between the selection test and each criterion measure was significant at the .05 level, with uncorrected r's of .47 for training score, .53 for training score with Spanish score removed, and .21 for Spanish language score. Although the validity coefficients obtained were significant, they were underestimates of the true relationships between the predictor and criteria. One factor affecting the accuracy of the obtained estimates was the restricted range of the sample of applicants.

Conclusions and Discussion. The results of this two-part validity study indicate that the selection test presently used with the law enforcement occupation of concern was valid for the final training score and that this validity was increased by the addition of five points for Spanish language ability. This increase in validity was apparently attributable to the Spanish language segment of the training program, since the correlation between the selection test and the Spanish language score increased while the correlation between the selection test and the final training score,with the Spanish score deleted,decreased with the addition of Spanish points to the selection test score. This was a reasonable and acceptable situation, in view of the fact that Spanish is a crucial part of this occupation.

It is important, in reporting the increased validity with the addition of language points, to emphasize the importance of Spanish language ability for this occupation. The importance of language should be taken into account before considering the addition of language points to selection test scores for other occupations. A similar increase in validity with the addition of language points would not be anticipated with an occupation where language plays a less than crucial role.

Rating Systems Are Not All Created Equal: The Baltimore Police Sergeant Experience

Robert G. Wendland, Deputy Personnel Director, Baltimore City Civil Service Commission

Introduction. Within the last five years, the City of Baltimore has developed, used and evaluated several types of supervisory rating systems, especially with its police and fire department promotion examinations. The impetus for this search for an effective rating system was a decision in federal court that the City had been guilty of employment discrimination in its Police Sergeant testing procedures since 1972. This paper reports experience with three types of supervisory ratings used in the Police Sergeant promotion procedures, with reference to the ratings used in 1978, 1981, and 1982. Supervisory ratings have been used in conjunction with a multiple choice written test, oral examinations, and seniority. It was felt that improved supervisory ratings with adequate variance, weighted more heavily in the promotion process might partly offset the adverse impact of the multiple choice test.

The 1978 Supervisory Rating System. The supervisory rating system used in 1978 and prior years was known as a Promotional Potential Rating. The rating was a one-scale global rating of the candidate's promotional potential made by the candidate's immediate supervisor and approved by the candidate's second-level supervisor. The scale ranged from 0 to 15; 12 was labeled "average." Written comments were required for ratings of 15 (outstanding) or less than 6 (not recommended). Little or no training was given on how or why to complete the rating form.

Supervisors naturally compared their candidates to those of other supervisors, and kept giving ratings at least as high as those given by other supervisors. Supervisors want to be nice and fair, and thus leniency was rampant. At least 60% of the candidates were rated "outstanding"; 88% were rated "outstanding" or "excellent" (rating 14). Even with this extreme leniency, a greater percentage of blacks were given numerical ratings of 13 and a lower percentage were given ratings of 15. The mean ratings for the blacks in 1978 was 14.31 with a standard deviation of .76. The mean rating for the whites was 14.64 with a standard deviation of .81.

The 1981 Rating System. In 1981 a consultant proposed supervisory rating system for 17 abilities or traits were used. Ratings for 12 of the 17 factors were to be "satisfactory/unsatisfactory"; the remaining 5 ratings were based on a forced "normalized ranking" system.

The "normalized ranking" system transforms a ranked list of candidates into a decile- or percentile-based list of candidates, assuming the candidates would be normally distributed. The process changes an ordinal rating into a percentile, based on the assumption that the candidate would fall at the midpoint of an interval under the normal curve divided into equal areas for the number of candidates.

Complicating the process were the number of candidates rated by each supervisor. Many supervisors had only one candidate to appraise; some supervisors had over 30 candidates to appraise. Thus while the candidates could be equitably evaluated on the pass/fail factors, some method of mathematically adjusting the scores would be needed for the ranking factors. How, for example, does a supervisor rank his candidates when there is only one candidate to rank? The consultant introduced 3 "hypothetical" candidates for each of the 5 ranking factors. Thus, each supervisor had at least 4 candidates to rank for each factor. Even this mechanism, however, failed to alleviate the problem that a supervisor who had only one candidate to appraise could only give a maximum rating of 78, while a supervisor with 10 or 20 candidates could give a score in the 90's. Most supervisors with only one candidate ranked their subordinates above all three hypothetical candidates. The ranking system required supervisors to break all ties. Thus, no two candidates could be assigned the same rank on those 5 factors which required ranking.

The results were predictable. For the first time in many years, the supervisors were forced to make hard, uncomfortable choices between their subordinates. Management perceived the inequities of ranking only 1 candidate and 3 hypothetical candidates; and ranking 20 candidates. When the results were published, the reaction was totally negative. Management was also upset at the ban on tied rankings, and was unprepared to explain rankings to the officers in the absence of some formal standards. Thus, the Police Department refused to reuse the 1981 normalized ranking system again in 1982.

The ranking factors accounted for almost all of the variance in the candidates' scores. Twelve of the seventeen factors were "satisfactory/unsatisfactory," and as true of most such rating systems, very few candidates are marked unsatisfactory. The plaintiffs charged that the ranking system allowed the influence of racial bias by the supervisors, without any real safeguards or checks and the white supervisors ranked black officers lower than white officers.

In 1982 the Commission developed a Behaviorally Anchored Rating (BAR) system to address the objections to the pass/fail and ranking factor system used in 1981. A BAR system uses written standards which objectively describe behaviors demonstrating different levels of the factor being measured. The system therefore, provides some uniform, explainable standards which would be applied equally to all candidates. Most articles on rating scales in the professional literature, furthermore, claim that BAR systems reduce rater biased.

The seven abilities to be measured by the supervisory appraisal in 1982 were:

1. Judgment in sizing up and evaluating situations.
2. Planning, organizing and scheduling ability.
3. Sensitivity and the ability to relate to others.
4. Report preparation and writing skills.
5. Ability to act under stress.
6. Oral communication skills.
7. Listening and attending skills.

To develop the behavioral statements measuring these abilities, the Commission held five meetings with police personnel with the rank of lieutenant and above. Each session was attended by five members of the Police Department and two Civil Service technicians familiar with behaviorally anchored systems. The police were given definitions of each factor to be rated. Based on these definitions, the police personnel provided objective, verifiable, behavioral statements which were pertinent to each factor. They were told to give three examples which they would consider "poor" and "good" actions. The behavioral statements for each factor were then given to approximately 60 police officials for ranking. After the statements were ranked from "best" to "worst," Civil Service technicians performed t-tests on the mean rankings for each statement and identified seven distinct levels of performance. The appraisal forms were printed containing the significant statements listed in rank order, but without numerical values. Training sessions were held including a videotaped presentation on how and why to complete the appraisal form, and a summary of the information contained in the rater's training manual.

The basic statistical parameters for the 1982 promotional appraisal are as follows:

Group	N	Mean	Standard Deviation
White	460	40.897	2.4120
Black	128	40.909	2.2008
Other	8	41.053	3.3661
Total	596	40.998	2.4041

Assuming raters were interchangeable, the reliability of this instrument was calculated at .23 for the total ratings. The major reason for the low reliability coefficient was the relatively low variance. Whereas the forced-ranking system used in 1981 artificially spread out the scores by assuming a normal distribution and by not allowing tied scores, the 1982 BAR system resulted in most candidates receiving a total score within a range of 2 points. The other major weakness of the 1982 system was in allowing raters to check a variable number of statements. The action was initiated by those candidates whose supervisors thought that more marks meant a higher score.

Unranked BAR Systems. The Commission has also tried BAR systems using statements in scrambled order. The numbers were too small to prove any significant reliability, but there was noticeable agreement when supervisors were instructed to mark only the 3 of the 8 or 9 statements which were most pertinent to the candidate. The scores were almost normally distributed and there were no significant racial effects. The major problem with unranked BAR systems is the insecurity of the raters. Management is still preoccupied with trying to ensure that the candidates they think best receive the highest scores.

Conclusions. Supervisory rating systems which are used for promotion decisions have very tumultuous effects on any organization, particularly large, decentralized organizations. Scales which rely only on the subjective judgments of the supervisors will fail if used with a large number of supervisors. Ranking systems are probably not appropriate in the public sector unless there are only a very few supervisors involved and all are ranking comparably small numbers of subordinates. Behaviorally anchored rating scales are very difficult to administer for job classes that are extremely diverse in their duties and assignments. Rating scales appear to be equally weak in reliability and internal consistency regardless of their design. Rating systems relying on different assumptions will differ markedly in their final results.

There is, of course, no answer as to what rating system is best. Rating systems are largely situation specific. Even today, because of the possible errors in all of the rating systems I have discussed, I cannot say which system is more valid. Thus far, the systems Baltimore has tried are unequal and unsatisfactory in one way or another, however, there are other techniques to be tried.

SYMPOSIUM

Implementing and Updating the Norfolk Police Department Performance Rating System

Chair: Robert J. Vance, Ohio State University
Discussant: Lt. H. P. Henson, Norfolk Police Department

Introduction. The Norfolk Police Department (NPD) Performance Rating System was designed to assist supervisors and commands in providing formal ratings and feedback to all police officers every six months. Officers at all levels and in all divisions were involved in the development process by writing and selecting the job-related statements used in the scales.

The purpose of the NPD Rating System was initiated to assist supervisors to formally identify and report the strengths and weaknesses in the performance of each of their subordinates. This information is used to provide formal feedback. Supervisors and subordinates focus on specific training needs, and thus increase job-related skills and productivity of subordinates. The ratings do not directly affect promotions, terminations, special assignments or other personnel decisions.

Within each dimension, specific and observable behaviors or activities written by Norfolk police officers,were selected as items for inclusion on the rating scales. Supervisors rate by marking the frequency with which they have observed each specific activity or behavior described on the scale. The rating scale is a five-point scale (5=Always; 1=Never).

Officers rated on the Patrol Officer Scale are also rated on a Forced-Choice Rating Scale. Forced-Choice Ratings are collected for research purposes only. They are not included as part of the Rating and Feedback System.

Once the rating data have been processed, Rating Feedback forms are generated and given to supervisors. Supervisors then schedule feedback sessions with each subordinate individually. The purpose of the session is to describe the ratings, to discuss performance strengths and weaknesses, to identify training needs, and to set goals for future performance improvement.

Perceptions of the NPD Rating System

E. Scott Wright, Ohio State University

The Behavioral Observation scales of the NPD rating system represent the content aspect of performance appraisal. More recently, researchers have expanded their view of issues which impact on performance appraisal and focused their attention on context and process issues as well. Representative of variables of interest in this expanded view are attitudinal correlates of rating system acceptability to users.

The present study represents an attempt to assess the reactions of raters and ratees in the Norfolk Police Department toward the rating system. We measured the perceptions of fairness and accuracy of performance ratings and rating system acceptability in samples of ratees. In addition, we obtained perceptions of rating system acceptability from samples of raters. We expected that performance ratings would be significantly related to both acceptability and perceived fairness and accuracy for ratees.

Participants. Ratee surveys were distributed on two occasions separated by a 9-month interval to a stratified sample of 155 members of the Norfolk Police Department. Rater surveys were distributed to a stratified sample of 30 sergeants on occasion one and 45 sergeants at time two.

Ratee Survey. The ratee survey consisted of 22 items which assessed reaction to the performance rating system on seven-point Likert scales with high values indicating positive attitudes. Dependent variables included perceived fairness and accuracy of ratings, measured by a single item and acceptability of the rating system, measured by a composite of items. Predictor variables included supervisory performance ratings, computed as the average of the 20 rating scale items, a five-item composite assessing perceptions of goal-setting, a three-item composite measuring feedback session atmosphere and eight single-item variables describing other reactions to the rating system.

Rater Survey Instruments. The rater survey consisted of 12 items, also measured on 7-point Likert scales. The dependent variable, acceptability of the performance rating system, was measured by a three-item composite. (Cronbach's alpha.83 for occasions one and two.) Predictors were nine single-item variables describing specific aspects of the rating system.

Analyses. For the ratee survey analyses, the dependent variable were regressed on the predictors in a hierarchical design. For the rater survey analysis stepwise regression was used to predict acceptance from the nine attitudinal variables.

Results: Ratee Survey. Contrary to our expectations, performance ratings did not contribute significantly to either occasion of attitudinal reaction to the rating system. Significant correlates of perceived fairness and accuracy at occasion one were the degree to which ratees received the ratings they expected and the opinion that supervisors had observed enough performance to rate them.

For the most part, different factors predicted rating system acceptance on occasion one. These were the occurrence of goal-setting, the adequacy of rating scale work dimensions- the degree to which ratees received the ratings they expected, and the degree to which adequate feedback was available on the job exclusive of that provided by the rating system.

The results for ratee system acceptance at occasion two substantially replicated those of occasion one. The adequacy of the rating scale work dimensions, the occurrence of goal-setting, and receiving the ratings expected were significant predictors of system acceptance.

Results: Rater Survey. For both occasions, rater attitudes toward the importance of encouraging discussion emerged as a predictor acceptance. A second predictor also emerged on both occasions; however, for occasion one the predictor was the extent to which the rating scale items were job-related, and for occasion two, the predictor was the usefulness of the supervisor's manual.

Discussion. The results concurred with findings of previous research to a great extent concerning predictors of perceived fairness and accuracy and performance rating system acceptability.

Rating system acceptability was determined by process, content, and context factors. The process variable, occurrence of goal-setting, and the content variable, adequacy and job-relatedness of the rating dimensions, were the most important contributors in both occasions.

For ratees, two factors contributed substantial amounts of variance explained in rating system acceptance on both occasions. Raters who felt that it was important to encourage subordinates to discuss ratings during feedback sessions were more accepting of the system. The other important factors in rater acceptance were content variables.

Process, content, and context factors are all important correlates of perceptions of rating fairness and accuracy, and of acceptance. The correlational nature of the data preclude casual inferences, but the results do suggest that the designer of developmentally oriented rating systems should ensure that rating scales are job related, provide open performance feedback on a regular basis in order to establish accurate expectations among ratees, tie goal setting to performance and feedback to past goals, encourage supervisors to attend as closely as possible to the job performance of subordinates, and convince supervisors of the importance of regular feedback.

Effects of Rater Training on Feedback and Goalsetting Behavior and Rating Errors: A Longitudinal Field Study

Peter S. Winne and Eduardo Salas, Old Dominion University

One of the most recalcitrant problems in organizational performance appraisal programs is the recurrent and ubiquitous problems of errors in supervisory ratings. Although a variety of rating errors have been identified or described in the literature, leniency and halo have received the most systematic attention. Leniency can be defined as the tendency of raters to assign ratings that are systematically higher or lower than the "true performance" of ratees. Halo refers to "a rater's failure to discrimate among conceptually distinct and potentially independent aspects of a ratee's behavior".

This research addresses two questions regarding error in ratings: the consistency and source of rating errors; and the results of a training experiment in which raters received training on feedback and goalsetting techniques.

Consistency and Source of Rating Errors. To investigate the consistency of rating errors across time, and the source of these errors, measures of leniency and halo were correlated between pairs of rating periods across five administrations of the rating scale within a three and one-half year period.

The research was conducted in a large metropolitan police department with approximately 550 uniformed personnel. The data were collected as part of an ongoing performance rating system which had been developed by Bernardin, Morgan and Winne in 1978. The sole purpose of the system is to provide formal feedback for employee development through supervisory ratings, which are completed approximately every 10 months.

In the study, leniency was computed as the mean of the 20 items of the rating instrument for ratees and as the grand mean of ratings for each rater. Halo was computed as the variance across the 20 items for each ratee and as the mean of within-scale variances across subordinates for each rater. Higher mean ratings indicated greater leniency effects. Higher variance indicated less halo.

For rater-based analyses, the intercorrelations for leniency demonstrated considerable consistency across time when the ratees were the same, but not when ratees were different. Since the variance between rating periods is apparently reliable only when the ratees were the same, the results imply that the source of leniency is a ratee effect. The rater-based halo analysis was indeterminate regarding both the consistency and locus of halo. In the leniency analysis all of the same-rater correlations were higher in magnitude than their respective different-rater correlations. These results suggest that leniency could be attributable both to raters and ratees. In the halo analyses, all of the adjacent period analyses were significant in the same-rater analysis while none of the correlations obtained significance in the different-rater analysis. These results suggest that halo, like leniency, is consistent across time and is primarily a rater effect.

Rater Feedback and Goal Setting Training. As indicated, the results of the longitudinal investigation of rating errors indicated that (1) they were reliable effects across considerable periods of time and (2) the source of these errors were primarily raters. If we can assume that leniency and halo in ratings can reduce their utility, a not unreasonable assumption, the question becomes "what kinds of interventions can we make to reduce these effects." Among the various methods are: (1) develop and use rating formats which reduce errors, (2) select as raters those personnel who are found to be resistent to making errors, (3) change the rating process itself with more frequent ratings and diary-keeping to reduce forgetting and distortion, (4) investigate the organizational determinants which enhance or reduce rating errors, (5) train supervisors directly to reduce errors by informing them of the various kinds and causes of these errors and ways to reduce them, (6) train supervisory skills in activities in the rating process which might indirectly impact on rating errors. For example, as we did here, goalsetting skills might be trained.

In the present study we tried to resturcture the performance appraisal environment through emphasis of goalsetting activities and to train raters to use more effective feedback and goalsetting techniques. Formal goal-setting activities between supervisors and subordinates were instituted by coordinating them with already established feedback policies in performance appraisals. The supervisors were given training in feedback and goalsetting through workshop, role playing, and lecture and discussion sessions. The performance appraisal process was modified to include formal goal setting activities between supervisors and their subordinates.

The survey instrument consisted of questions on feedback and goalsetting which were included as part of larger questionnaires for raters and ratees soliciting information about overall perceptions of the performance rating system. On the ratee questionnaire, 9 items about the feedback session and goalsetting activities were asked at each of the three administrations. The items were grouped into four categories about different aspects of the session. These measures were (1) utility of the session; (2) atmospheres; (3) goalsetting activities; and (4) feedback activities. On the rater questionnaire, two items relevant to provision of and important of the feedback session were asked.

Description of the Training. The training was designed to provide information and techniques about feedback and goalsetting to supervisors that would aid them during the feedback session. The effects of the training were examined through analyses of ratee and rater questionnaire data and the measures of rating errors. The results suggest that the main effect of the intervention was on goalsetting but that feedback effectiveness was not modified substantially.

Examination of the consistency and potential source of rating errors across five rating periods indicated that both leniency and halo exhibited substantial reliable variance across time which was primarily attributable to rater tendencies rather than to differences between ratees when raters were held constant. Thus, the typical assumption that leniency and halo represent stable rater traits was supported in this study. In the goalsetting training and intervention, subordinates perceived changes in the conduct of the feedback session, with more goalsetting activities, but these changes did not have a large effect on the perceptions of the utility of the feedback session in improving behavior.

SYMPOSIUM

The Administration of a Psychological Testing Program in Police and Correctional Agencies

Chair: Robert E. Inwald, Hilton Research, Inc.

Role of Inwald Personality Inventory and Minnesota Multiphasic Personality Inventory as Predictors of Correction Officers Job Performance by Race

Elizabeth J. Shusman and Robin E. Inwald, Hilton Research, Inc.

Unfortunately, there is inadequate information available on racial differences among candidates or correction officers on either personality tests or actual job performance measures. This report involves studies of two personality inventories. The Minnesota Multiphasic Personality Inventory (MMPI) and the IPI (Inwald Personality Inventory) in their utilization as predictors of correction official job performance.

Due to the lack of information regarding racial differences in psychological testing of correction officers, the following research questions were posed: (1) can two personality measures (the MMPI and the IPI) successfully predict performance for different racial groups?; (2) do these two tests have different predictive abilities for whites, blacks, and Hispanics?

Subjects. The subjects of this study consisted of 716 male urban correction officers who completed application and hiring procedures between 1980 and 1982. Of the officers who continued to work at the law enforcement agency, 45% were white (n=298), 34% were black (n=225), 19% were Hispanic (n=127), and the 2% balance indicated other racial group membership. Twenty-one percent of the terminated officers were white (n=11), 61% were black (n=31) and 18% were Hispanic (n=9).

Method. Applicants for employment in a large urban correction department completed a battery of questionnaires and psychological inventories which included a "Personal History Questionnaire," a source of biographical data, the MMPI, and the IPI.

In order to measure previously unassessed behavioral patterns that might be more directly related to subsequent law enforcement job performance, the IPI was developed for use in this program. It was designed specifically for use with law enforcement officers and many items were developed directly from candidate interviews. Validation and cross-validation studies using correction officers and police officers indicated that on several job-related criteria, the IPI consistently predicted actual performance for a greater number of officers than did the MMPI.

In order to measure job performance for those officers subsequently hired, records were kept for each recruit by a supervising officer regarding termination from the department, absence, lateness and disciplinary interviews received during a ten-month probationary period.

Results. In the development of separate prediction equations for each of the four criterion measures (termination, absence, lateness, and disciplinary interviews), the direct method of discrimination function analysis was employed. All 26 IPI scales and 14 MMPI scales, including the standard clinical and validity scales, and one experimental scale, "MacAndrews Alcohol," were used.

With respect to 309 white male recruits studied, 298 remained on the job (herein referred to as "successes") and 11 (4%) were terminated ("failures"). A preliminary prediction equation formed from the IPI alone accurately classified 88% of the recruits. The equation developed for whites from the MMPI had a prediciton accuracy of 74%, a result not statistically significant. One MMPI scale, "Hypochondriasis," had a significantly higher mean for "failures." Together, the IPI and MMPI scales correctly identified 91% of the whites as to "success" or "failure" on the job. Other analyses were presented.

With few exceptions, analyses of race differences on the three actual job performance criteria (absence, lateness, and disciplinary interviews) did not reveal statistically significant results. These findings are notable, however, because of the high rate of accuracy in predicting officers who exhibited "positive" or "negative" on-job behaviors. In general, for each race,the IPI correctly classified a greater number of officers than did the MMPI. The only exception to this pattern was on one criterion, "lateness" for blacks and for whites, where the MMPI classified a greater number of officers. When both inventories were used together to develop prediction equations, accuracy increased for all job criteria measured.

The authors' complete paper presents detailed statistical analyses.

Discussion. Two trends were apparent when the prediction rates of the IPI and MMPI were compared for each race on the individual criteria. First, the IPI identified a greater percentage of officers with regard to their on-job behaviors than did the MMPI. This was true on all criteria examined with the exception of "lateness" for whites and for blacks. Secondly, when all scales of the IPI and MMPI were combined, prediction accuracy was at its highest. This research suggests that the IPI and MMPI can successfully predict job performance for different racial groups. There appeared to be little difference on overall prediction rates for the white, black and Hispanic male recruits studied. Thus, these tests may be largely non-discriminatory measures, useful as pre-employment screening tools in law enforcement agencies.

A Validation and Cross-Validation Study of Correction Officer Job Performance as Predicted by the IPI and MMPI

Elizabeth J. Shusman, Robin E. Inwald and Beth Landa, Hilton Research, Inc.

The predictive validity of two psychological inventories, the Minnesota Multiphasic Personality Inventory (MMPI) and the Inwald Personality Inventory (IPI) was examined for 716 male correction officer recruits for retention or termination as well as incidence of absence, lateness and formal disciplinary interviews. A cross-validation analysis of the three on-job performance criteria on a "cross-validation" sample of 265 officers was then conducted.

Predictive validity studies comparing the IPI to the MMPI on objective as well as subjective performance criteria (such as the incidence of absence, lateness, and formal disciplinary procedures) have been conducted with 596 correction officers (Inwald and Shusman) and 329 police officers (Inwald and Shusman). In these studies, prediction equations were developed employing test scale scores as the independent variables in discriminant function analyses. Equations were computed for the IPI scales alone, the standard MMPI scales, and for the IPI in combination with the MMPI. The IPI alone consistently predicted a greater number of correction and police officers as to the presence or absence of "negative" job behavior (ranging from 60% to 82% of the cases) than did the MMPI alone (from 54% to 71%) on the different criteria. Together, the IPI and MMPI accurately identified the greatest number of individuals. Inwald and Sakales report that the IPI and MMPI used together, using a sample of 175 correction officers, correctly predicted the presence or absence of "negative" job behavior for 69% (on the "Absence" criterion) to 77% ("Lateness") of the officers.

Study I. The subjects consisted of 716 male urban correction officer recruits. Of these, 665 remained in the department and 51 were terminated from the agency within ten months after the hiring date. Several observed differences were noted between the terminated and retained officers. A comparison of biographical data showed that, prior to the application date: (1) thirty-three percent of the terminated officers versus 22% of the retained officers were arrested at least once; (2) almost five times the number of terminated officers than retained officers had been in some form of trouble while in the military (29% versus 6%), for those who had military experience (61% of the terminated and 31% of the retained officers); (3) twice the number of terminated officers than retained officers had collected welfare payments since the age of 18 (12% versus 6%); and (4) twenty-four percent of the terminated officers versus 15% of the retained officers had previously been fired from at least one job.

For Study II (cross-validation) study, the 665 officers remaining in the department were subsequently randomly divided into two groups for a cross-validation analysis. The first, "analysis" group consisted of 400 (60%) and the second, "cross-validation" group numbered 265 (40%). Overall, the two groups appeared to have similar characteristics.

Method: Studies I and II. As applicants, all subjects were required to complete the MMPI, IPI, and a "Personal History Questionnaire," from which biographical data were gleaned. After hiring, recruits entered a training academy and were placed on a ten-month probationary period. At no time were supervising officers allowed to view the psychological ratings of a recruit. During probation, records were kept by the supervising officers on the incidence of absence, lateness, corrective interviews and disciplinary actions for each individual. At the completion of ten months, the department reserved the right to terminate any recruit from employment.

Both Study I and Study II (cross-validation study) involve an examination of the ability of the IPI alone, or the MMPI alone, and of IPI in conjunction with the MMPI to predict job performance. Study I involved an examination of retention - termination rates for all 665 retained officers and 51 terminated officers. Study II involved only actual job performance of retired officers in terms of the criteria of absence, lateness, and disciplinary interviews.

Discriminant function analyses using the Discriminant Function subprogram of the Statistical Package for the Social Sciences (SPSS) were performed. All 26 scales of the IPI were entered into the prediction equations as were the 13 clinical scales and one experimental scale, "MacAndrews Alcohol," coded "low" (1), "moderate" (2), and "high" (3), of the MMPI. Equations were developed from the 26 IPI scales alone, from the 14 MMPI scales alone, as well as from all 40 IPI and MMPI scales.

Very detailed statistical tables accompanied the presentation of this report, supporting the results (available upon request).